

1. Conductors, Semiconductors and Diodes

1. [Simple Conduction](#)
2. [Introduction to Semiconductors](#)
3. [Doped Semiconductors](#)
4. [P-N Junction: Part I](#)
5. [P-N Junction: Part II](#)
6. [Gauss' Law](#)
7. [Depletion Width](#)
8. [Forward Biased PN Junctions](#)
9. [The Diode Equation](#)
10. [Reverse Biased/Breakdown](#)
11. [Diffusion](#)
12. [Light Emitting Diode](#)
13. [LASER](#)
14. [Solar Cells](#)

2. Bipolar Transistors

1. [Introduction to Bipolar Transistors](#)
2. [Transistor Equations](#)
3. [Transistor I-V Characteristics](#)
4. [Common Emitter Models](#)
5. [Small Signal Models](#)
6. [Small Signal Model for Bipolar Transistor](#)

3. FETs

1. [Introduction to MOSFETs](#)
2. [Basic MOS Structure](#)
3. [Threshold Voltage](#)
4. [MOS Transistor](#)
5. [MOS Regimes](#)
6. [Plotting MOS I-V](#)
7. [Models](#)
8. [Inverters and Logic](#)

9. [Transistor Loads for Inverters](#)
10. [CMOS Logic](#)
11. [JFET](#)
12. [Electrostatic Discharge and Latch-Up](#)
4. IC Manufacturing
 1. [Introduction to IC Manufacturing Technology](#)
 2. [Silicon Growth](#)
 3. [Doping](#)
 4. [Fick's First Law](#)
 5. [Fick's Second Law](#)
 6. [Photolithography](#)
 7. [Integrated Circuit Well and Gate Creation](#)
 8. [Applying Metal/Sputtering](#)
 9. [Integrated Circuit Manufacturing: Bird's Eye View](#)
 10. [Diffused Resistor](#)
 11. [Yield](#)
5. Introduction to Transmission Lines
 1. [Distributed Parameters](#)
 2. [Telegrapher's Equations](#)
 3. [Transmission Line Equation](#)
 4. [Transmission Line Examples](#)
 5. [Exciting a Line](#)
 6. [Terminated Lines](#)
 7. [Bounce Diagrams](#)
 8. [Cascaded Lines](#)
6. AC Steady-State Transmission
 1. [Review of Phasors](#)
 2. [A/C Line Behavior](#)
 3. [Terminated Lines](#)
 4. [Line Impedance](#)
 5. [Crank Diagram](#)
 6. [Standing Waves/VSWR](#)

7. [Bilinear Transform](#)
8. [The Smith Chart](#)
9. [Introduction to Using the Smith Chart](#)
10. [Simple Calculations with the Smith Chart](#)
11. [Power](#)
12. [Finding the Load Impedance](#)
13. [Matching](#)
14. [Introduction to Parallel Matching](#)
15. [Single Stub Matching](#)
16. [Double Stub Matching](#)
17. [Odds and Ends](#)

Simple Conduction

Our initial studies will more or less be a review of topics in electricity that you may have seen before in physics. However, if experience is any guide, there is no great harm in going back over this material, for it seems that for many students, the whole concept of just how electricity actually works is just a little hazy. Considering that you hope to be called an electrical engineer one of these days, this might even be a good thing to know!

Most of the "laws" of how electricity behaves are really just mathematical representations of a number of empirical observations, based on some assumptions and guesses which were made in attempt to bring the "laws" into a coherent whole. Early investigators (Faraday, Gauss, Coulomb, Henry etc....all those guys) determined certain things about this strange "invisible" thing called electricity. In fact, the electron itself was only discovered a little over 100 years ago. Even before the electron itself was observed, people knew that there were two kinds of electric charge, which were called **positive** and **negative**. Like charges exhibit a repulsive force between them and opposite charges attract one another. This force is proportional to the product of the absolute value of positive and negative charge, and varies inversely with the square of the distance between them. Different charge carriers have different mass, some are very light, and others are significantly heavier. Electrical charges can experience forces, and can move about. Since force times distance equals work, a whole system of energy (**potential** as well as **kinetic**) and energy loss had to be described. This has lead to our current system of electrostatics and electrodynamics, which we will not review now but bring up along the way as things are needed.

Just to make sure everyone is on the same footing however, let's define a few quantities now, and then we will see how they interact with one another as we go along.

The total charge in some region is defined by the symbol Q and it has units of Coulombs. The fundamental unit of charge (that of an electron or a proton) is symbolized either by a little q or by e . Since we'll use e for other things, in this course we will try to stick with q . The **charge of an electron**, q , has a value of 1.6×10^{-19} Coulombs.

Since charge can be distributed throughout a region with varying concentrations, we will also talk about the **charge density**, $\rho(\nu)$, which has units of $\frac{\text{Coulombs}}{\text{cm}^3}$. (In this book, we will use a modified MKS system of units. In keeping with most workers in the solid-state device field, volume will usually be expressed as a cubic centimeter, rather than a cubic meter - a cubic meter of silicon is just far too much!) In most cases, the charge density is not uniform but is a function of where we are in space. Thus, when we have $\rho(\nu)$ distributed throughout some volume, V

Equation:

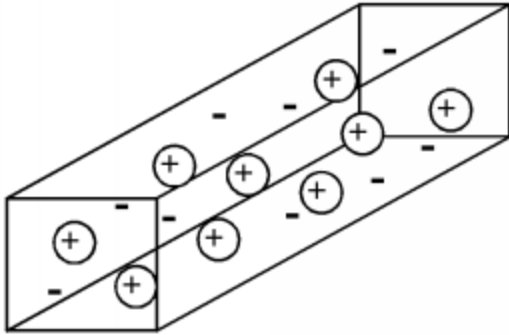
$$Q = \int \rho(\nu) \, d\nu$$

describes the total charge in that volume.

We know that when we apply an electric field to a charge that there is a force exerted on it, and that if the charge is able to move it will do so. The motion of charge gives rise to an **electric current**, which we call I . The current is a measure of how much charge is passing a given point per unit time ($\frac{\text{Coulombs}}{\text{second}}$).

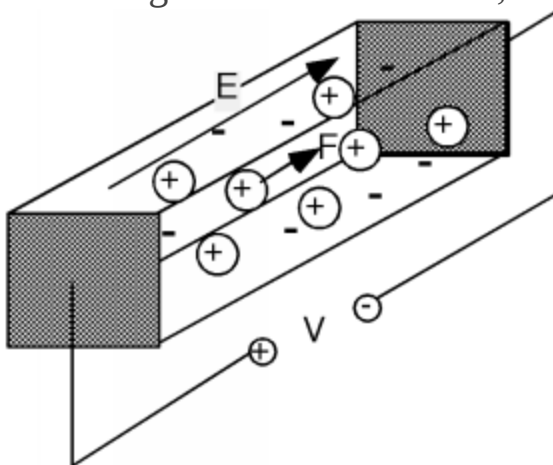
It will be helpful if we have some kind of model of how electricity flows in a conductor. There are several approaches which one can take, some more intuitive than others. The one we will look at, while not correct in the strictest sense, still gives a very good picture of how electrical conduction works, and is perfectly fine to use in a variety of situations. In the **Drude theory** of conduction, the initial hypothesis consists of a solid, which contains mobile charges which are free to move about under the influence of an applied electric field. There are also fixed charges of polarity opposite that of the mobile charges, so that everywhere within the solid, the net charge density is zero. (This hypothesis is based on the model of the atom, with a positively charged nucleus and negatively charged electrons surrounding it. In a solid, the atoms are fixed in position in the lattice, but it is assumed that some of the electrons can break free of their "host" atom and move about to other places within the solid.) In our model, let us choose the polarity of the mobile charges to be positive; this is not usually

the case, but we can avoid a lot of "minus ones" this way, and have a better chance of ending up with the right answer in the end.



Model of a conductor.

As shown in [\[link\]](#), the model of the conductor consists of a number of mobile positive charges (represented by the balls with the "+" sign in them) and an equal number of fixed negative charges (represented by the bare "-" sign). In subsequent figures, we will leave out the fixed charge, since it can not contribute in any way to the conduction process, but keep in mind that it is there, and that the total net charge is zero within the material. Each of the mobile charge carriers has a mass, m , and an amount of charge, q .



Applying a potential to a conductor

In order to have some conduction, we have to apply a potential or voltage across the sample ([\[link\]](#)). We do this with a battery, which creates a potential difference, V , between one end of the sample and the other. We will make the simplest assumption that we can, and say that the voltage, V , gives rise to a uniform electric field within the sample. The magnitude of the electric field is given simply by

Equation:

$$E = \frac{V}{L}$$

where L is the length of the sample, and V is the voltage which is placed across it. (In truth, we should be showing E as well as subsequent forces etc. as vectors in our equations, but since their direction will be obvious, and unambiguous, let's keep things simple, and just write them as scalars.)

Electric potential, or voltage, is just a measure of the change in potential energy per unit charge going from one place to another. Since energy, or work is simply force times distance, if we divide the energy per unit charge by the distance over which that potential exists, we will end up with force per unit charge, or electric field, E . If you are not sure about what you just read, write it out as equations, and see that it is so.

The electric field will exert a force on the movable charges (And the fixed ones too for that matter, but since they can not go anywhere, nothing happens to them). The force is given simply as the product of the electric field strength times the charge

Equation:

$$F = qE$$

The force acts on the charges and causes them to accelerate according to Newton's equations of motion

Equation:

$$\begin{aligned} F &= m \frac{dv(t)}{dt} \\ &= qE \end{aligned}$$

or

Equation:

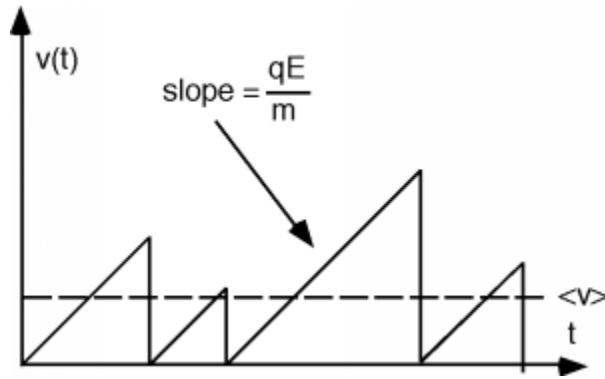
$$\frac{d}{dt} v(t) = \frac{qE}{m}$$

Thus, the velocity of a particle with no initial velocity will increase linearly with time as:

Equation:

$$v(t) = \frac{qE}{m} t$$

The rate of acceleration is proportional to the strength of the electric field, and inversely proportional to the mass of the particle. The particle can not continue to accelerate forever however. Since it is located within a solid, sooner or later it will collide with either another carrier, or perhaps one of the fixed atoms within the solid. We will assume that the collision is completely inelastic, and that after a collision, the particle comes to a stop, only to be accelerated again by the electric field. If we were to make a plot of the particles velocity as a function of time, it might look something like [\[link\]](#).



Velocity as a function of time
for charge carrier

Although the particle achieves various velocities, depending upon how much time there is between collisions, there will be some average velocity, \bar{v} , which will depend upon the details of the collision process. Let us define a scattering time τ_s which will give us that average velocity when we multiply it times the acceleration of the particle. That is:

Equation:

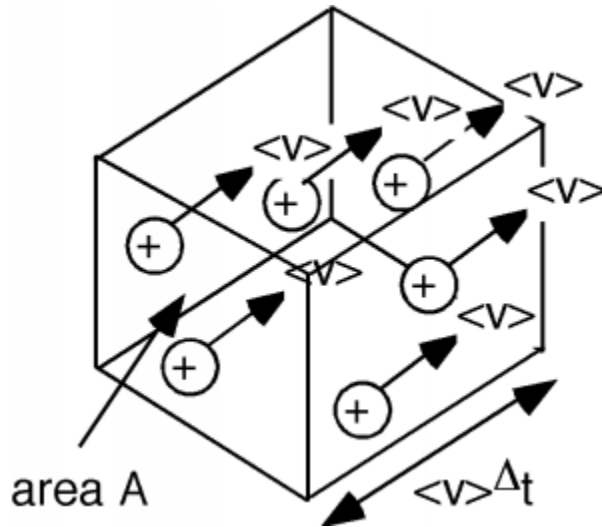
$$\bar{v} = \frac{qE\tau_s}{m}$$

or

Equation:

$$\tau_s \equiv \frac{m\bar{v}}{qE}$$

Now let's take a look at just a small section of the conductor ([\[link\]](#)). It will have the cross section of the sample, A , but will only be $\bar{v}\Delta(t)$ long, where $\Delta(t)$ is just some arbitrary time interval.



Section of the conductor

After a time $\Delta(t)$ has passed, all of the charges within the box will have left it, as they are all moving with the same average velocity, \bar{v} . If the density of charge carriers in the conductor is n per unit volume, then the number of carriers N within our little box is just n times the volume of the box

$$\bar{v}\Delta(t)A$$

Equation:

$$N = n\bar{v}\Delta(t)A$$

Thus the total charge, Q , which leaves the box in time $\Delta(t)$ is just qN . The current flow, I , is just the amount of charge which flows out of the box per unit time

Equation:

$$\begin{aligned}
 I &= \frac{qn\bar{v}\Delta(t)A}{\Delta(t)} \\
 &= qn\bar{v}A \\
 &= \frac{q^2n\tau_s EA}{m} \\
 &= \frac{Q}{\Delta(t)}
 \end{aligned}$$

We now have two choices, we can look at our result from a field quantity point of view, in which case we will be interested in the **current density**, J , which is just the current, I , divided by the cross-sectional area

Equation:

$$\begin{aligned}
 J &= \frac{I}{A} \\
 &= \frac{q^2n\tau_s}{m} E \\
 &= \sigma E
 \end{aligned}$$

where σ is called the **conductivity** of the material. If we look at the conductor from a macroscopic point of view, then we are interested in the relationship between the voltage and the current. The voltage is just the electric field times the length of the sample, and the current is just the current density times its cross sectional area. Thus we have

Equation:

$$\begin{aligned}
 I &= AJ \\
 &= A\sigma E \\
 &= A\sigma \frac{V}{L}
 \end{aligned}$$

or

Equation:

$$\begin{aligned}
 V &= \frac{L}{\sigma A} I \\
 &= RI
 \end{aligned}$$

where R is the resistance of the sample. We have discovered **Ohm's law**!

Note that [\[link\]](#) tells us that the resistance of the sample is proportional to its length (the longer the sample, the higher the resistance) and inversely proportional to its cross sectional area (the fatter the sample, the lower the resistance). The sample resistance is also inversely proportional to the conductivity σ of the sample. Sometimes, instead of conductivity, the **resistivity**, ρ , is specified for a resistive material. The resistivity is simply the inverse of the conductivity

Equation:

$$\sigma = \frac{1}{\rho}$$

and thus:

Equation:

$$R = \frac{\rho L}{A}$$

And, in an effort towards completeness, there is one other quantity which you might run into, and that is the carrier **mobility**, μ . The mobility is just the proportionality factor between the average velocity of the particle and the electric field. That is:

Equation:

$$\bar{v} = \mu E$$

You should check that the following two relationships are correct:

Equation:

$$\sigma = nq\mu$$

Equation:

$$\mu = \frac{q\tau_s}{m}$$

If we take an ordinary conductor (and we will have to define later what we mean by that) and heat it up, the atoms within the material start to vibrate faster due to the elevated temperature, and the carriers suffer significantly more collisions. The mean collision time τ_s decreases, and hence the conductivity goes down, and the resistance of the sample goes up.

Introduction to Semiconductors

If we only had to worry about simple conductors, life would not be very complicated, but on the other hand we wouldn't be able to make computers, CD players, cell phones, i-Pods and a lot of other things which we have found to be useful. We will now move on, and talk about another class of conductors called semiconductors.

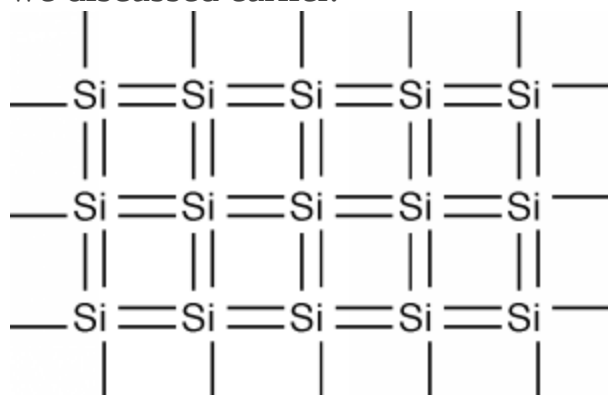
In order to understand semiconductors and in fact to get a more accurate picture of how metals, or normal conductors actually work, we really have to resort to quantum mechanics. Electrons in a solid are very tiny objects, and it turns out that when things get small enough, they no longer exactly following the classical "Newtonian" laws of physics that we are all familiar with from everyday experience. It is not the purpose of this course to teach you quantum mechanics, so what we are going to do instead is describe the results which come from looking at the behavior of electrons in a solid from a quantum mechanical point of view.

Solids (at least the ones we will be talking about, and especially semiconductors) are crystalline materials, which means that they have their atoms arranged in an ordered fashion. We can take silicon (the most important semiconductor) as an example. Silicon is a group IV element, which means it has four electrons in its outer or valence shell. Silicon crystallizes in a structure called the **diamond** crystal lattice. This is shown in [\[link\]](#). Each silicon atom has four covalent bonds, arranged in a tetrahedral formation about the atom center.



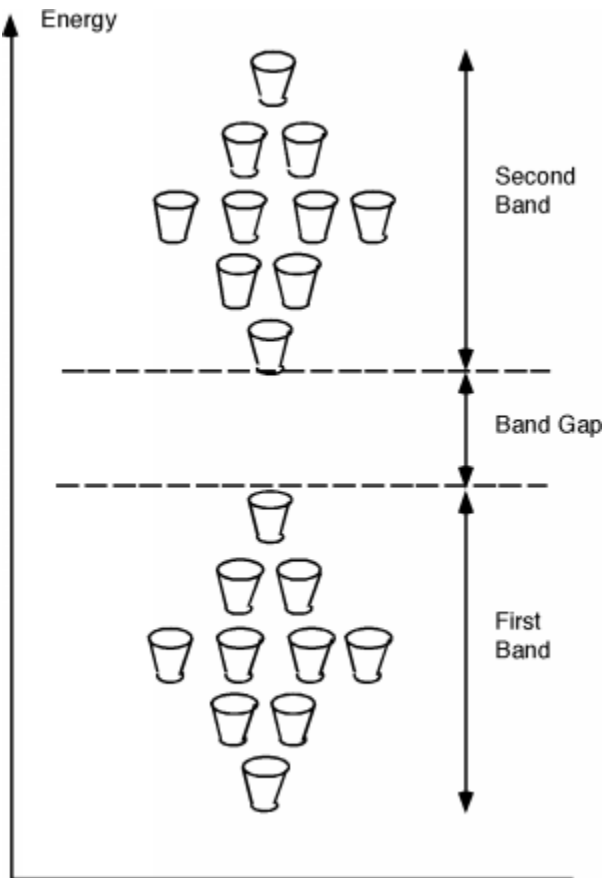
Crystal structure of silicon

In two dimensions, we can schematically represent a piece of single-crystal silicon as shown in [\[link\]](#). Each silicon atom shares its four valence electrons with valence electrons from four nearest neighbors, filling the shell to 8 electrons, and forming a stable, periodic structure. Once the atoms have been arranged like this, the outer valence electrons are no longer strongly bound to the host atom. The outer shells of all of the atoms blend together and form what is called a **band**. The electrons are now free to move about within this band, and this can lead to electrical conductivity as we discussed earlier.



A 2-D representation of a silicon crystal

This is not the complete story however, for it turns out that due to quantum mechanical effects, there is not just one band which holds electrons, but several of them. What will follow is a very qualitative picture of how the electrons are distributed when they are in a periodic solid, and there are necessarily some details which we will be forced to gloss over. On the other hand this will give you a pretty good picture of what is going on, and may enable you to have some understanding of how a semiconductor really works. Electrons are not only distributed throughout the solid crystal spatially, but they also have a distribution in energy as well. The potential energy function within the solid is periodic in nature. This potential function comes from the positively charged atomic nuclei which are arranged in the crystal in a regular array. A detailed analysis of how electron **wave functions**, the mathematical abstraction which one must use to describe how small quantum mechanical objects behave when they are in a periodic potential, gives rise to an energy distribution somewhat like that shown in [\[link\]](#).



Schematic of the first two bands
in a periodic solid showing
energy levels and bands

Firstly, unlike the case for free electrons, in a periodic solid, electrons are not free to take on any energy value they wish. They are forced into specific energy levels called **allowed states** which are represented by the cups in the figure. The allowed states are not distributed uniformly in energy either. They are grouped into specific configurations called **energy bands**. There are no allowed levels at zero energy and for some distance above that. Moving up from zero energy, we then encounter the first energy band. At the bottom of the band there are very few allowed states, but as we move up in energy, the number of allowed states first increases, and then falls off again. We then come to a region with no allowed states, called an energy **band gap**. Above the band gap, another band of allowed states exists. This

goes on and on, with any given material having many such bands and band gaps. This situation is shown schematically in [\[link\]](#), where the small cups represent allowed energy levels, and the vertical axis represents electron energy.

It turns out that each band has exactly $2N$ allowed states in it, where N is the total number of atoms in the particular crystal sample we are talking about. (Since there are 10 cups in each band in the figure, it must represent a crystal with just 5 atoms in it. Not a very big crystal at all!) Into these bands we must now distribute all of the valence electrons associated with the atoms, with the restriction that we can **only put one electron into each allowed state**. (This is the result of something called the **Pauli exclusion principle**.) Since in the case of silicon there are 4 valence electrons per atom, we would **just** fill up the first two bands, and the next would be empty. (If we make the logical assumption that the electrons will fill in the levels with the lowest energy first, and only go into higher lying levels if the ones below are already filled.) This situation is shown in [\[link\]](#).

Here, we have represented electrons as small black balls with a "-" sign on them. Indeed, the first two bands are completely full, and the next is empty. What will happen if we apply an electric field to the sample of silicon? Remember the diagram we have at hand right now is an **energy** based one, we are showing how the electrons are distributed in energy, not how they are arranged spatially. On this diagram we can not show how they will move about, but only how they will change their energy as a result of the applied field. The electric field will exert a force on the electrons and attempt to accelerate them. If the electrons are accelerated, then they must increase their kinetic energy. Unfortunately, there are no empty allowed states in either of the filled bands. An electron would have to jump all the way up into the next (empty) band in order to take on more energy. In silicon, the gap between the top of the highest most occupied band and the lowest unoccupied band is 1.1 eV. (One eV is the potential energy gained by an electron moving across an electrical potential of one volt.) The **mean free path** or distance over which an electron would normally move before it suffers a collision is only a few hundred angstroms ($\approx (300 \times 10^{-8})$ cm) and so you would need a very large electric field (several hundred thousand $\frac{\text{volts}}{\text{cm}}$) in order for the electron to pick up enough energy to "jump the gap".

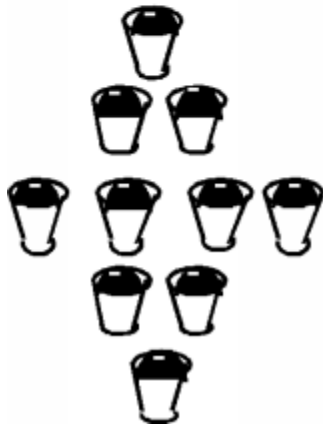
This makes it appear that silicon would be a very bad conductor of electricity, and in fact, very pure silicon is very poor electrical conductor.



Silicon,
with first
two
bands full
and the
next
empty

A metal is an element with an **odd** number of valence electrons so that a metal ends up with an upper band which is just half full of electrons. This is

illustrated in [\[link\]](#). Here we see that one band is full, and the next is just half full. This would be the situation for the Group III element aluminum for instance. If we apply an electric field to these carriers, those near the top of the distribution can indeed move into higher energy levels by acquiring some kinetic energy of motion, and easily move from one place to the next. In reality, the whole situation is a bit more complex than we have shown here, but this is not too far from how it actually works.



Electron
distribution for
a metal or good
conductor

So, back to our silicon sample. If there are no places for electrons to "move" into, then how does silicon work as a "semiconductor"? Well, in the first place, it turns out that not all of the electrons are in the bottom two bands. In silicon, unlike say quartz or diamond, the band gap between the top-most full band, the next empty one is not so large. As we mentioned above it is only about 1.1 eV. So long as the silicon is not at absolute zero temperature, some electrons near the top of the full band can acquire enough thermal energy that they can "hop" the gap, and end up in the upper band, called the **conduction band**. This situation is shown in [\[link\]](#).



Thermal excitation
of electrons across
the band gap

In silicon at room temperature, roughly 10^{10} electrons per cubic centimeter are thermally excited across the band-gap at any one time. It should be noted that the excitation process is a continuous one. Electrons are being excited across the band, but then they fall back down into empty spots in the lower band. On average however, the 10^{10} in each cm^3 of silicon is what you will find at any given instant. Now 10 billion electrons per cubic centimeter **seems** like a lot of electrons, but lets do a simple calculation. The mobility of electrons in silicon is about $1000 \frac{\text{cm}^2}{\text{volt-sec}}$. Remember, mobility times electric field yields the average velocity of the carriers. Electric field has units of $\frac{\text{volts}}{\text{cm}}$, so with these units we get velocity in $\frac{\text{cm}}{\text{sec}}$ as we should.) The charge on an electron is 1.6×10^{-19} coulombs. Thus from [this equation](#):

Equation:

$$\begin{aligned}\sigma &= nq\mu \\ &= 10^{10} (1.6 \times 10^{-19}) 1000 \\ &= 1.6 \times 10^{-6} \frac{\text{mhos}}{\text{cm}}\end{aligned}$$

If we have a sample of silicon 1 cm long by (1 mm) (1 mm) square, it would have a resistance of

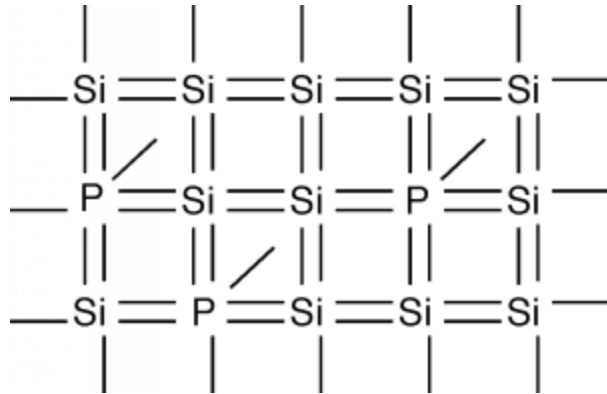
Equation:

$$\begin{aligned}R &= \frac{L}{\sigma A} \\ &= \frac{1}{(1.6 \times 10^{-6}) 0.1^2} \\ &= 62.5 \text{M}\Omega\end{aligned}$$

which does not make it much of a "conductor". In fact, if this were all there was to the silicon story, we could pack up and move on, because at **any** reasonable temperature, silicon would conduct electricity very poorly.

Doped Semiconductors

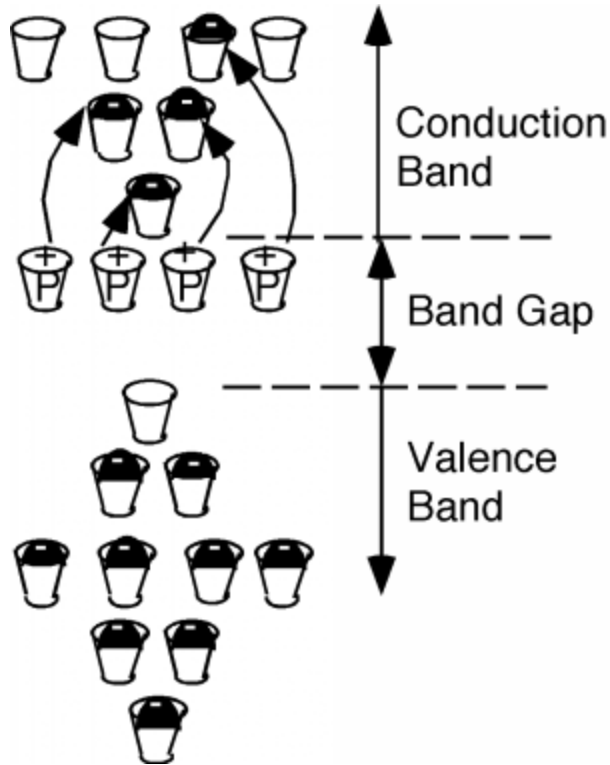
To see how we can make silicon a useful electronic material, we will have to go back to its crystal structure. Suppose somehow (and we will talk about how this is done later) we could substitute a few atoms of phosphorus for some of the silicon atoms.



A silicon crystal "doped" with phosphorus

If you sneak a look at the periodic table, you will see that phosphorus is a group V element, as compared with silicon which is a group IV element. What this means is the phosphorus atom has **five** outer or **valence** electrons, instead of the four which silicon has. In a lattice composed mainly of silicon, the extra electron associated with the phosphorus atom has no "mating" electron with which it can complete a shell, and so is left loosely dangling to the phosphorus atom, with relatively low binding energy. In fact, with the addition of just a little thermal energy (from the natural or latent heat of the crystal lattice) this electron can break free and be left to wander around the silicon atom freely. In our "energy band" picture, we have something like what we see in [\[link\]](#). The phosphorus atoms are represented by the added cups with P's on them. They are new allowed energy levels which are formed within the "band gap" near the bottom of the first empty band. They are located close enough to the empty (or "conduction") band, so that the electrons which they contain are easily excited up into the conduction band. There, they are free to move about and contribute to the electrical conductivity of the sample. Note also, however,

that since the electron has left the vicinity of the phosphorus atom, there is now one more proton than there are electrons at the atom, and hence it has a net positive charge of $1q$. We have represented this by putting a little "+" sign in each P-cup. Note that this positive charge is fixed at the site of the phosphorous atom called a **donor** since it "donates" an electron up into the conduction band, and is not free to move about in the crystal.



Silicon doped with phosphorus

How many phosphorus atoms do we need to significantly change the resistance of our silicon? Suppose we wanted our 1 mm by 1 mm square sample to have a resistance of one ohm as opposed to more than $60\text{ M}\Omega$. Turning the resistance equation around we get

Equation:

$$\begin{aligned}
 \sigma &= \frac{L}{RA} \\
 &= \frac{1\Omega}{1 \times 0.1^2} \\
 &= 100 \frac{\text{mho}}{\text{cm}}
 \end{aligned}$$

And hence (If we continue to assume an electron mobility of $1000 \frac{\text{cm}^2}{\text{volt sec}}$)

Equation:

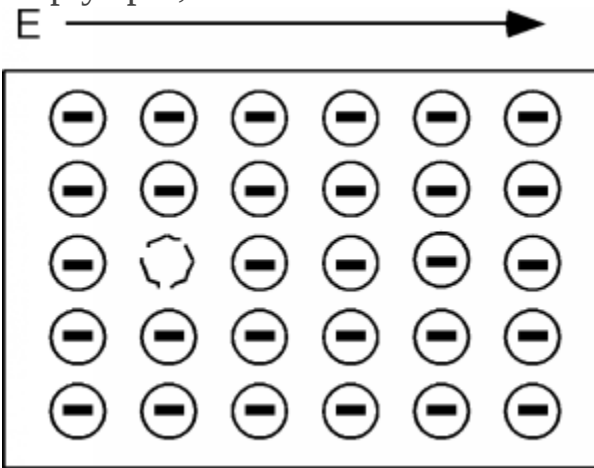
$$\begin{aligned}
 n &= \frac{\sigma}{q\mu} \\
 &= \frac{100}{(1.6 \times 10^{-19})1000} \\
 &= 6.25 \times 10^{17} \text{cm}^3
 \end{aligned}$$

Now adding more than 6×10^{17} phosphorus atoms per cubic centimeter might seem like a lot of phosphorus, until you realize that there are almost 10^{24} silicon atoms in a cubic centimeter and hence only one in every 1.6 million silicon atoms has to be changed into a phosphorus one to reduce the resistance of the sample from several 10s of $\text{M}\Omega$ down to only one Ω . This is the real power of semiconductors. You can make dramatic changes in their electrical properties by the addition of only minute amounts of impurities. This process is called "**doping**" the semiconductor. It is also one of the great challenges of the semiconductor manufacturing industry, for it is necessary to maintain fantastic levels of control of the impurities in the material in order to predict and control their electrical properties.

Again, if this were the end of the story, we still would not have any calculators, stereos or "Agent of Doom" video games (Or at least they would be very big and cumbersome and unreliable, because they would have to work using vacuum tubes!). We now have to focus on the few "empty" spots in the lower, almost full band (Called the **valence band**.) We will take another view of this band, from a somewhat different perspective. I must confess at this point that what I am giving you is even further from the way things really work, then the "cups at different energies" picture we have been using so far. The problem is, that in order to do things right, we have to get involved in momentum phase-space, a lot more quantum

mechanics, and generally a bunch of math and concepts we don't really need in order to have some idea of how semiconductor devices work. What follow below is really intended as a motivation, so that you will have some feeling that what we state as results, is actually reasonable.

Consider [\[link\]](#). Here we show all of the electrons in the valence, or almost full band, and for simplicity show one missing electron. Let's apply an electric field, as shown by the arrow in the figure. The field will try to move the (negatively charged) electrons to the left, but since the band is almost completely full, the only one that can move is the one right next to the empty spot, or **hole** as it is called.

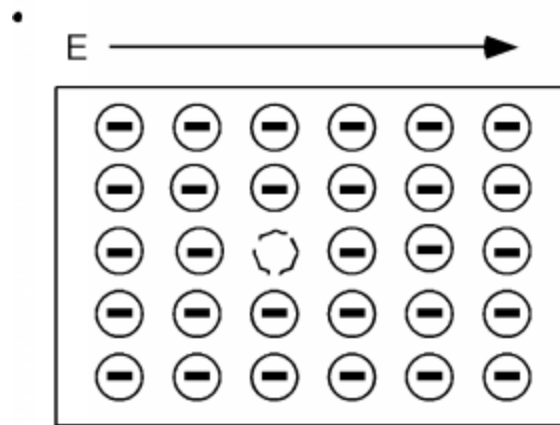


Band full of electrons, with one missing

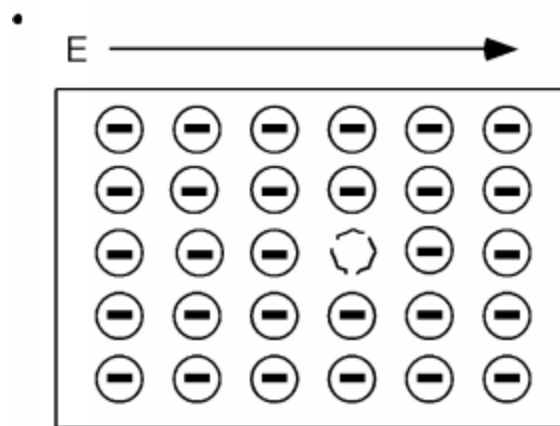
One thing you may be worrying about is what happens to the electrons at the ends of the sample. This is one of the places where we are getting a somewhat distorted view of things, because we should really be looking in momentum, or wave-vector space rather than "real" space. In that picture, they magically drop off one side and "reappear" on the other. This doesn't happen in real space of course, so there is no easy way we can deal with it.

A short time after we apply the electric field we have the situation shown in [\[link\]](#), and a little while after that we have [\[link\]](#). We can interpret this motion in two ways. One is that we have a net flow of negative charge to

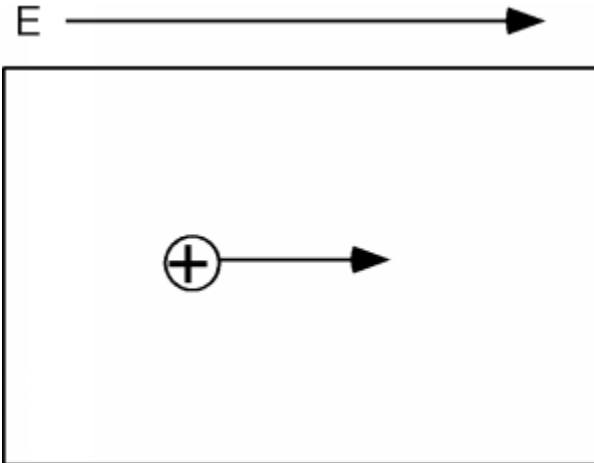
the left, or if we consider the effect of the aggregate of all the electrons in the band (which we have to do because of quantum mechanical considerations beyond the scope of this book) we could picture what is going on as a single positive charge, moving to the right. This is shown in [\[link\]](#). Note that in either view we have the same net effect in the way the total **net** charge is transported through the sample. In the mostly negative charge picture, we have a net flow of negative charge to the left. In the single positive charge picture, we have a net flow of positive charge to the right. Both give the same sign for the current!



Motion of the "missing" electron with an electric field



Further motion of the "missing electron" spot



Motion of a "hole" due to an applied electric field

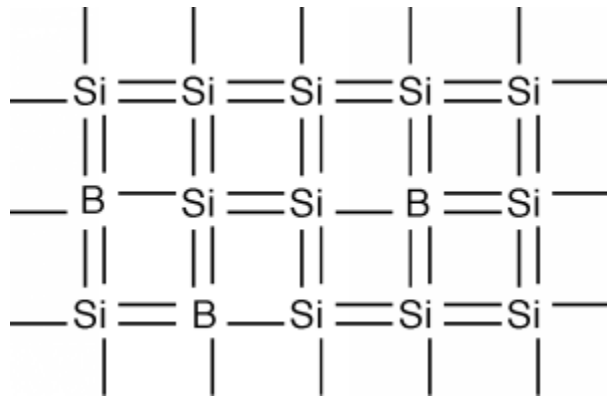
Thus, it turns out, we can consider the consequences of the empty spaces moving through the co-ordinated motion of electrons in an almost full band as being the motion of positive charges, moving wherever these empty spaces happen to be. We call these charge carriers "holes" and they too can add to the total conduction of electricity in a semiconductor. Using ρ to represent the density (in cm^{-3}) of spaces in the valence band and μ_e and μ_h to represent the mobility of electrons and holes respectively (they are usually not the same) we can modify [this equation](#) to give the conductivity σ , when both electrons' **holes** are present.

Equation:

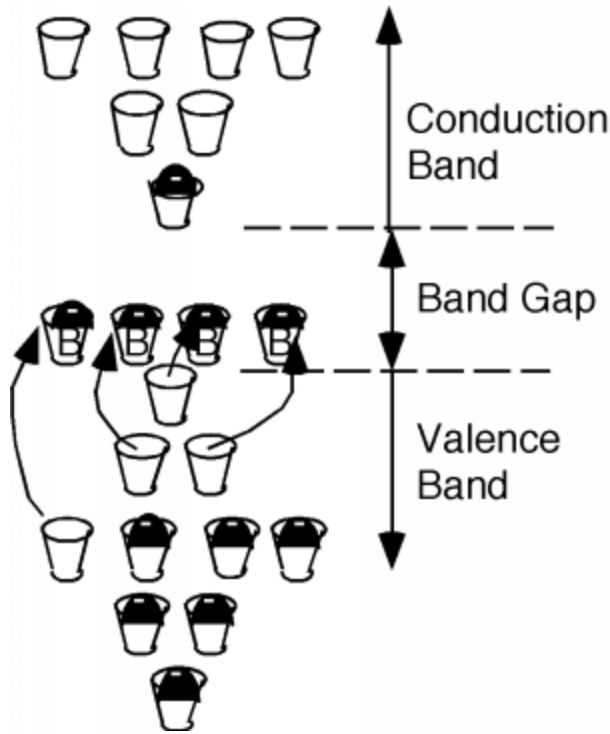
$$\sigma = nq\mu_e + \rho q\mu_h$$

How can we get a sample of semiconductor with a **lot** of holes in it? What if, instead of phosphorus, we dope our silicon sample with a group III element, say boron? This is shown in [\[link\]](#). Now we have some **missing**

orbitals, or places where electrons could go if they were around. This modifies our energy picture as follows in [\[link\]](#). Now we see a set of new levels introduced by the boron atoms. They are located within the band gap, just a little way above the top of the almost full, or valence band. Electrons in the valence band can be thermally excited up into these new allowed levels, creating empty states, or holes, in the valence band. The excited electrons are stuck at the boron atom sites called **acceptors**, since they "accept" an electron from the valence band, and hence act as **fixed** negative charges, localized there. A semiconductor which is doped predominantly with acceptors is called **p-type**, and most of the electrical conduction takes place through the motion of holes. A semiconductor which is doped with donors is called **n-type**, and conduction takes place mainly through the motion of electrons.



Silicon doped with Boron



P-type silicon, due to boron acceptors

In n-type material, we can assume that all of the phosphorous atoms, or **donors**, are fully ionized when they are present in the silicon structure. Since the number of donors is usually much greater than the native, or intrinsic electron concentration, ($\approx (10^{10}\text{cm}^{-3})$), if N_d is the density of donors in the material, then n , the electron concentration, $\approx (N_d)$.

If an electron deficient material such as boron is present, then the material is called **p-type** silicon, and the hole concentration is just $p \simeq N_a$ the concentration of **acceptors**, since these atoms "accept" electrons from the valence band.

If both donors and acceptors are in the material, then whichever one has the higher concentration wins out. (This is called **compensation**.) If there are more donors than acceptors then the material is n-type and $n \simeq N_d - N_a$. If there are more acceptors than donors then the material is p-type and

$p \simeq N_a - N_d$. It should be noted that in most compensated material, one type of impurity usually has a much greater (several order of magnitude) concentration than the other, and so the subtraction process described above usually does not change things very much. ($10^{18} - 10^{16} \simeq 10^{18}$).

One other fact which you might find useful is that, again, because of quantum mechanics, it turns out that the **product** of the electron and hole concentration in a material must remain a constant. In silicon at room temperature:

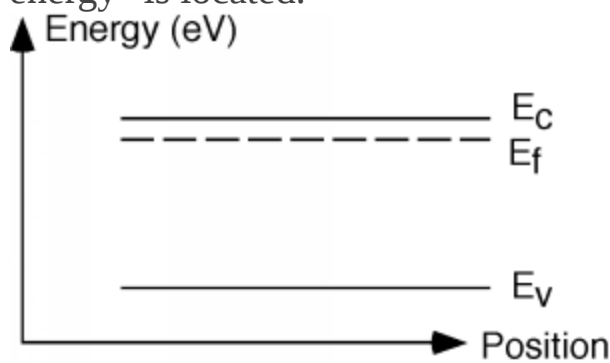
Equation:

$$np \equiv n_i^2 \simeq 10^{20} \text{cm}^{-3}$$

Thus, if we have an n-type sample of silicon doped with 10^{17} donors per cubic centimeter, then n , the electron concentration is just and p , the hole concentration, is $\frac{10^{20}}{10^{17}} = 10^3 \text{cm}^{-3}$. The carriers which dominate a material are called **majority carriers**, which would be the electrons in the above example. The other carriers are called **minority carriers** (the holes in the example) and while 10^3 might not seem like much compared to 10^{17} the presence of minority carriers is still quite important and can not be ignored. Note that if the material is undoped, then it must be that $n = p$ and $n = p = 10^{10}$.

The picture of "cups" of different allowed energy levels is useful for gaining a pictorial understanding of what is going on in a semiconductor, but becomes somewhat awkward when you want to start looking at devices which are made up of both n and p type silicon. Thus, we will introduce one more way of describing what is going on in our material. The picture shown in [\[link\]](#) is called a band diagram. A **band diagram** is just a representation of the energy as a function of position with a semiconductor device. In a band diagram, positive energy for electrons is upward, while for holes, positive energy is downwards. That is, if an electron moves **upward**, its potential energy **increases** just as a with a normal mass in a gravitational field. Also, just as a mass will "fall down" if given a chance, an electron will move down a slope shown in a band diagram. On the other hand, holes gain energy by moving **downward** and so they have a tendency to "float"

upward if given the chance - much like a bubble in a liquid. The line labeled E_c in [\[link\]](#) shows the edge of the conduction band, or the bottom of the lowest unoccupied allowed band, while E_v is the top edge of the valence, or highest occupied band. The band gap, E_g for the material is obviously $E_c - E_v$. The dotted line labeled E_f is called the **Fermi level** and it tells us something about the chemical equilibrium energy of the material, and also something about the type and number of carriers in the material. More on this later. Note that there is no zero energy level on a diagram such as this. We often use either the Fermi level or one or other of the band edges as a reference level on lieu of knowing exactly where "zero energy" is located.



An electron band-diagram for a semiconductor

The distance (in energy) between the Fermi level and either E_c and E_v gives us information concerning the density of electrons and holes in that region of the semiconductor material. The details, once again, will have to be begged off on grounds of mathematical complexity. (Take Semiconductor Devices (ELEC 462) in your senior year and find out how it really works!) It turns out that you can say:

Equation:

$$n = N_c e^{-\frac{E_c - E_f}{kT}}$$

Equation:

$$p = N_v e^{-\frac{E_f - E_v}{kT}}$$

Both N_c and N_v are constants that depend on the material you are talking about, but are typically on the order of 10^{19}cm^{-3} . The expression in the denominator of the exponential is just Boltzman's constant, k , times the temperature T of the material (in absolute temperature or Kelvin).

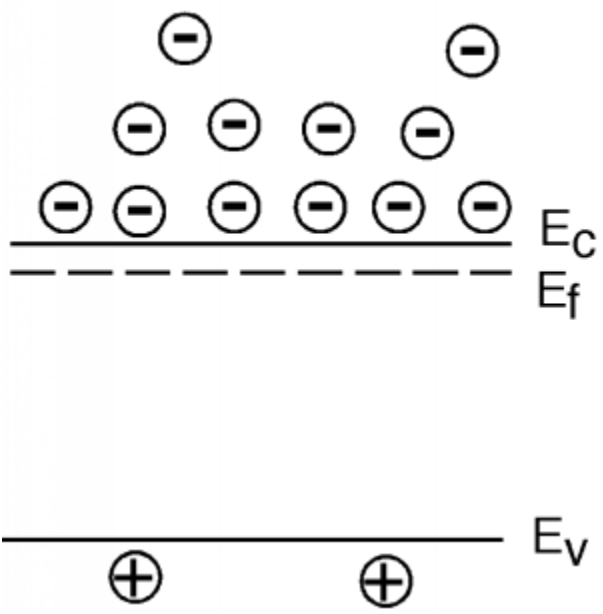
Boltzman's constant $k = (8.63 \times 10^{-5}) \frac{\text{eV}}{\text{K}}$. At room temperature $kT = 1/40$ of an electron volt. Look carefully at the numerators in the exponential. Note first that there is a minus sign in front, which means the bigger the number in the exponent, the fewer carriers we have. Thus, the top expression says that if we have n-type material, then E_f must not be too far away from the conduction band, while if we have p-type material, then the Fermi level, E_f must be down close to the valence band. The closer E_f gets to E_c the more electrons we have. The closer E_f gets to E_v , the more holes we have. [\[link\]](#) therefore must be for a sample of n-type material. Note also that if we know how heavily a sample is doped (That is, we know what N_d is for example) and from the fact that $n \simeq N_d$ we can use [\[link\]](#) to find out how far away the Fermi level is from the conduction band

Equation:

$$E_c - E_f = kT \ln \left(\frac{N_c}{N_d} \right)$$

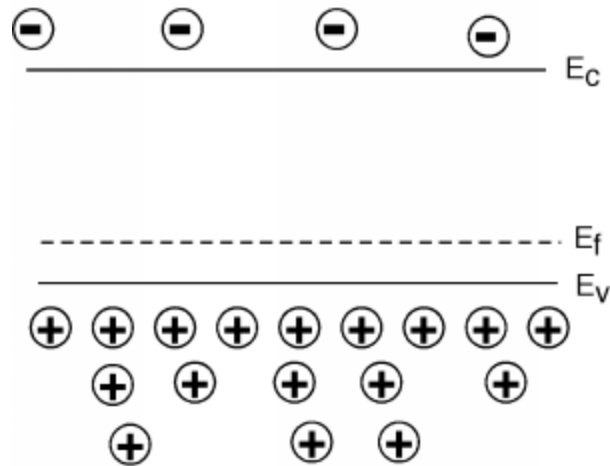
To help further in our ability to picture what is going on, we will often add to this band diagram, some small signed circles to indicate the presence of mobile electrons and holes in the material. Note that the electrons are spread out in energy. From our "cups" picture we know they like to stay in the lower energy states if possible, but some will be distributed into the higher levels as well. What is distorted here is the scale. The band-gap for silicon is 1.1 eV, while the **actual** spread of the electrons would probably only be a few tenths of an eV, not nearly as much as is shown in [\[link\]](#). Lets look at a sample of p-type material, just for comparison. Note that for holes, increasing energy goes **down** not up, so their distribution is inverted from that of the electrons. You can kind of think of holes as bubbles in a glass of soda or beer, they want to float to the top if they can. Note also for both n

and p-type material there are also a few "minority" carriers, or carriers of the opposite type, which arise from thermal generation across the band-gap.

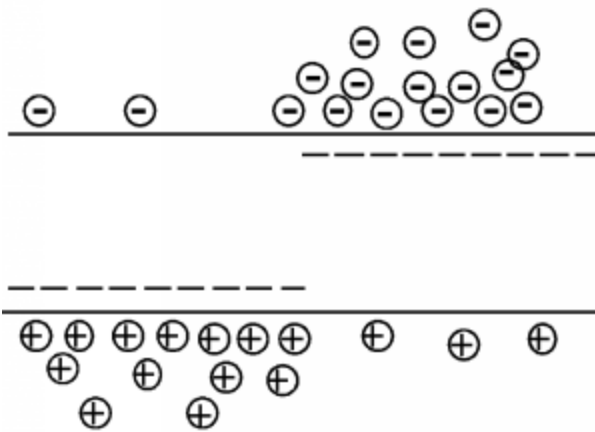


Band diagram for an n-type semiconductor

P-N Junction: Part I



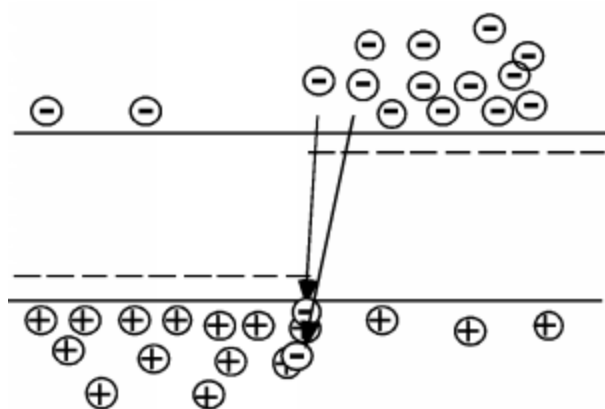
Band diagram for a p-type semiconductor



A non-equilibrium p-n junction

We are now ready to make an actual useful device! Let's take a piece of n-type material, and a piece of p-type material, and stick them together, as shown in [\[link\]](#). This way we will be making a **pn-junction**, or **diode**, which will be our first real electric device other than a simple resistor.

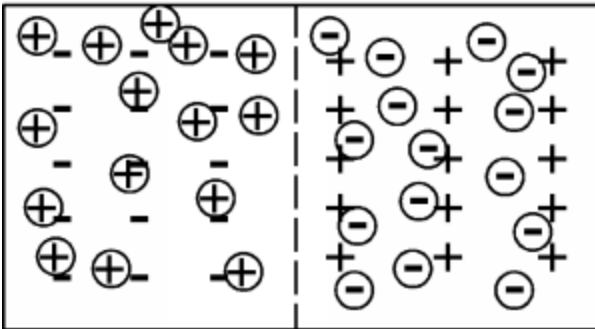
There are a couple of things wrong with [\[link\]](#). First of all, one of the rules regarding the Fermi level is that when you have a system at **equilibrium** (that is, when it is at rest, and is not being influenced by external forces such as thermal gradients, electrical potentials etc.), the Fermi level must be the same everywhere. Secondly, we have a big bunch of holes on the right and a big bunch of electrons on the left, and so we would expect, that in the absence of some force to keep them this way, they will start to spread out until their distribution is more or less equal everywhere. Finally, we remember that a hole is just an absence of an electron, and since an electron in the conduction band can lower the system energy by falling down into one of the empty hole states, it seems likely that this will happen. This process is called **recombination**. The place where this is most likely to occur, of course, would be right at the junction between the n and p regions. This is shown in [\[link\]](#).



Recombination of holes and electrons

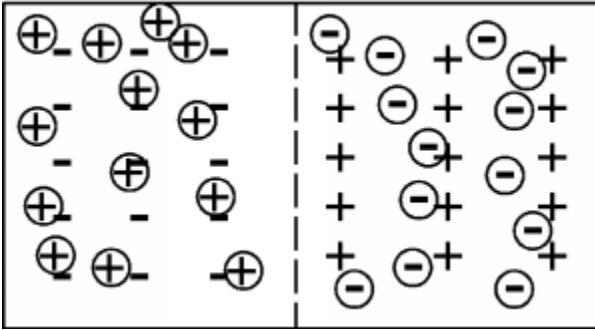
Now it might seem that this recombination effect might just go on and on, until there are no carriers left in the sample. This is not the case however. In order to see what brings everything to a halt, we need yet another diagram. [\[link\]](#) is more physical than what we have been looking at so far. It is a picture of the actual p-n junction, showing both the holes and the electrons. We also need to put in the donors and acceptors however, if we want to see what goes on. The fixed (can't move around) charges of the donors and acceptors are represented by simple "+" and "-" signs. They are arranged in

a nice lattice-like arrangement to remind us that they are stuck to the crystal lattice. (In reality however, even though they are stuck in the crystal lattice, there are so few of them compared to the silicon atoms that their distribution would be quite random.) For the mobile holes and electrons, we will stay with the little circles with charge signs in them. These are randomly distributed, to remind us that they are free to move about the crystal.



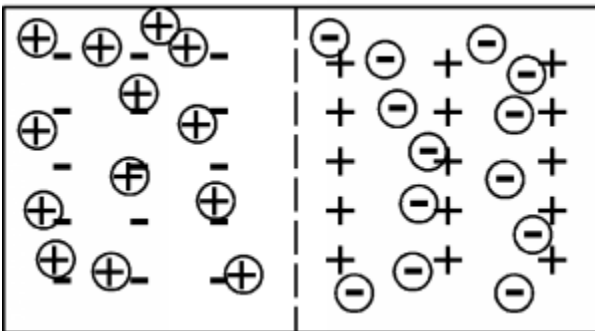
Spatial schematic of a p-n junction

We will now have to allow some of the holes and electrons (again near the junction) to recombine. Remember, when an electron and a hole recombine, they both are annihilated and disappear. Note that this process conserves charge (and if we could calculate it) momentum as well. There is obviously some energy lost, but this will simply show up as vibrations, or heat, within the crystal lattice. Or, in the case of an LED, as light emitted from the device. See, already we know enough about semiconductors to understand (somewhat) how an actual device works. Light coming from an LED is simply the energy which is released when an electron and hole recombine. We will take a look at this in more detail later. Let's allow some recombination to occur, as shown in [\[link\]](#).



The junction after some recombination has occurred

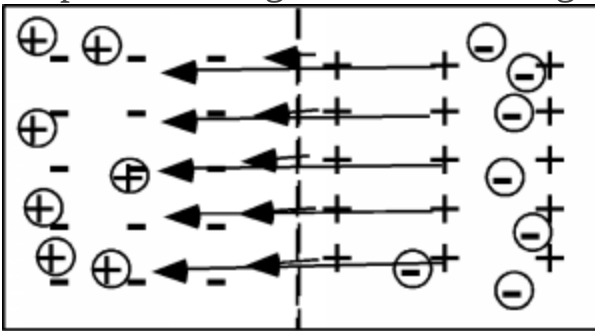
And then in [\[link\]](#) some more.....



After further recombination

P-N Junction: Part II

If you look closely at these pictures, you will notice something. As we remove more and more electrons and holes, we are starting to "uncover" the fixed charges associated with the donors and acceptors. We are making what is known as a **depletion region**, so named because it is **depleted** of mobile carriers (holes and electrons). The uncovered net charge in the depletion region is separated, with negative charge in the p-region, and positive charge in the n-region. What will such a charge separation give rise to? Why, an electric field! Of course! Which way will the field point? The electric field which arises from a separation of charges always goes from the positive charge, towards the negative charge. This is shown in [\[link\]](#).

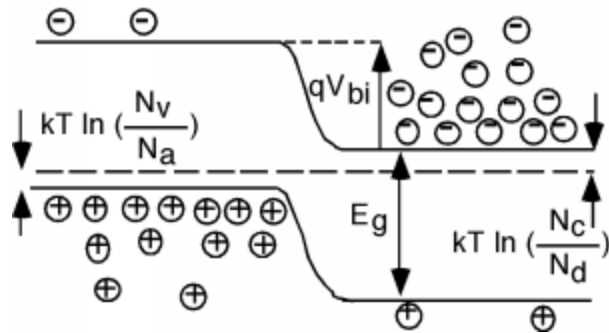


The pn-junction with the resultant built-in electric field

What effect will this field have on our device? It will have the tendency to push the holes back into the p-region and the electrons into the n-region. This is just what we need to counteract the recombination which has been going on, and hopefully bring it to a stop.

Now try to think through what effect this field could have on our energy band diagram. The band diagram is for electrons, so if an electron moves from the right hand side of the device (the n-region) towards the left hand side (the p-region), it will have to move through an electric field which is opposing its motion. This means it has to do some work, or in other words, the potential energy for the electron must go up. We can show this on the band diagram by simply shifting the bands on the left hand side upward, to

indicate that there is a shift in potential energy as electrons move from right to left across the junction.



Energy band diagram for a p-n junction at equilibrium

The shift of the bands, which is just the difference between the location of the Fermi level in the n-region and the Fermi level in the p-region, is called the **built-in potential**, V_{bi} . This built-in potential keeps the majority of holes in the p-region, and the electrons in the n-region. It provides a potential barrier, which prevents current flow across the junction. (On the band diagram we have to multiply the built-in potential V_{bi} by the charge of an electron, q , so that we can represent the shift in energy in terms of **electron volts**, the unit of potential energy used in band diagrams.)

How big is V_{bi} ? This is not too hard to figure out. Let's look at [\[link\]](#) a little more carefully. Remember, we know from [this equation](#) and [this equation](#) that since $n = N_d$ in the n-region and $p = N_a$ in the p-region, we can relate the distance of the Fermi level from E_c and E_f by

Equation:

$$E_c - E_f = kT \ln\left(\frac{N_c}{N_d}\right)$$

and

Equation:

$$E_f - E_v = kT \ln\left(\frac{N_v}{N_a}\right)$$

Look at [\[link\]](#) and see if you can agree that

Equation:

$$\begin{aligned} qV_{\text{BI}} &= E_g - (E_c - E_f) - (E_f - E_v) \\ &= E_g - kT \ln\left(\frac{N_c}{N_d}\right) - kT \ln\left(\frac{N_v}{N_a}\right) \\ &= E_g - kT \ln\left(\frac{N_c N_v}{N_d N_a}\right) \end{aligned}$$

Where N_d and N_a are the doping densities in the n and p side respectively. Remember, $kT = 1/40 \text{ eV} = 0.025 \text{ eV}$, $E_g = 1.1 \text{ eV}$ and N_c and N_v are both $\approx (10^{19})$. Thus,

$$qV_{\text{BI}} = 1.1 \text{ eV} - 0.025 \text{ eV} \ln\left(\frac{10^{38}}{N_d N_a}\right)$$

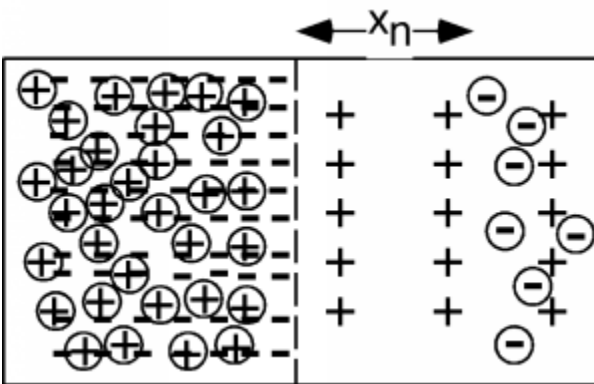
Here the q in front of the V_{BI} and the e in eV are both the charge of 1 electron and they cancel out making

$$V_{\text{BI}} = \left(1.1 - 0.025 \ln\left(\frac{10^{38}}{N_d N_a}\right) \right) \text{ volts}$$

Suppose both N_d and N_a are both about [10 to the 15th] - not uncommon values. How big would the built-in potential be in this case?

It turns out that we can actually derive some specific details about the depletion region if we make only a coupled of simplifying (and often justified) assumptions. In order to make the math easier, and also because many p-n junctions are built this way, we will consider what is known as a **one sided junction**. [\[link\]](#) is a picture of such a beast: In this diode, one side is much more heavily doped than the other. In this particular example, the p-side is heavily doped, and the n-side is relatively lightly doped. We

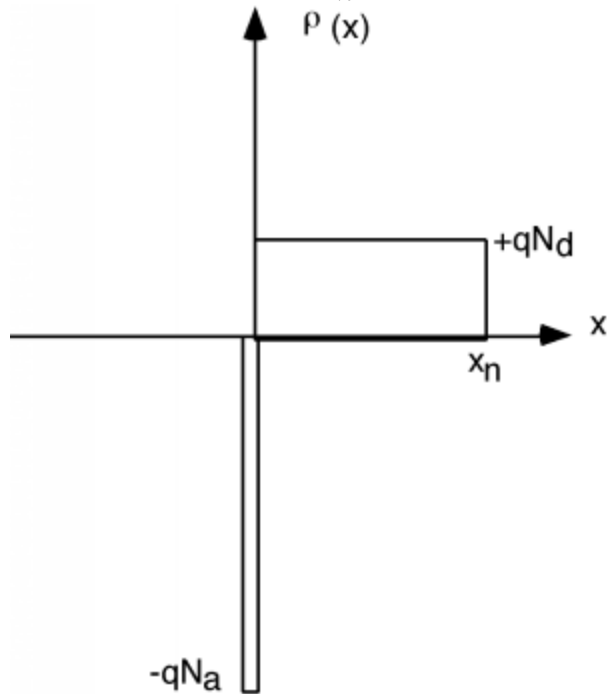
can not show the true picture here, because typically, the more heavily doped side will be doped **several orders of magnitude** greater than the lightly doped side. Typical values might be $N_a = 10^{19}$ and $N_d = 10^{16}$. Regardless of how big the difference is however, there must be exactly the same amount of "uncovered" charge on both side of the junction. Why? Because each time a hole and electron recombine to form the depletion region, they each leave behind either a donor or an acceptor. A careful count of the exposed charge in [\[link\]](#) shows that I was careful enough to draw my figure accurately for you. We do not need to have a one-sided diode to do the analysis that will follow, but the equations are easier to solve if we do.



An example of a one-sided diode

In order to proceed from here, the first thing we do is make a plot of the charge density $\rho(x)$ as we move through the junction. Naturally, in the bulk, since the holes and the acceptors (in the p-side), or the electrons and the donors (in the n-side) just equal one another, the net charge density is zero. In the depletion region, the charge density is $-(-q)N_a$ on the p-side and $(q)N_d$ on the donor side. (All the mobile carriers are gone, and we are left with just the charged acceptors or donors.) We will make the assumption that on the n-side, the depletion extends a distance $-x_n$ from the junction. On the p-side, the acceptor charge density is so large, that we will treat it as a δ -function, with essentially no width. The areas of the two boxes must be the same (equal amount of positive and negative charge) and

hence, the tall thin box actually has a width of $\frac{N_d}{N_a}x_n$, which, since N_a is several orders of magnitude greater than N_d , means that the tall box has a very very small width compared to the lower, wider one, which is qN_d tall, and has a width of x_n .



Charge density as a function of position

Gauss' Law

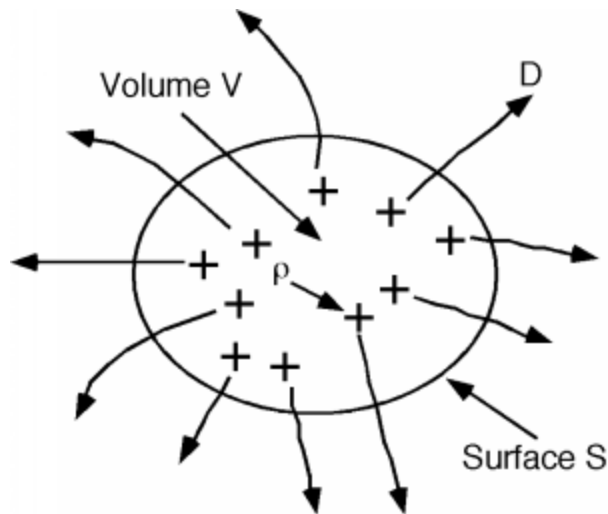
Now we have to review some field theory. We will be using fields from time to time in this course, and when we need some aspect of field theory, we will introduce what we need at that point. This seems to make more sense than spending several weeks talking about a lot of abstract theory without seeing how or why it can be useful.

The first thing we need to remember is **Gauss' Law**. Gauss' Law, like most of the fundamental laws of electromagnetism comes not from first principle, but rather from empirical observation and attempts to match experiments with some kind of self-consistent mathematical framework. Gauss' Law states that:

Equation:

$$\oint_{s,} D \, dS = Q_{\text{encl}}$$
$$= \oint_{v,} \rho(v) \, dV$$

where D is the **electric displacement vector**, which is related to the **electric field vector**, E , by the relationship $D = \epsilon E$. ϵ is called the **dielectric constant**. In silicon it has a value of $1.1 \times 10^{-12} \frac{F}{\text{cm}}$. (Note that D must have units of $\frac{\text{Coulombs}}{\text{cm}^2}$ to have everything work out OK.) Q_{encl} is the total amount of charge enclosed in the volume V , which is obtained by doing a volume integral of the charge density $\rho(v)$.



Pictorial representation of Gauss' Law.

[\[link\]](#) just says that if you add up the surface integral of the displacement vector D over a closed surface S , what you get is the sum of the total charge enclosed by that surface. Useful as it is, the integral form of Gauss' Law, (which is what [\[link\]](#) is) will not help us much in understanding the details of the depletion region. We will have to convert this equation to its differential form. We do this by first shrinking down the volume V until we can treat the charge density $\rho(v)$ as a constant ρ , and replace the volume integral with a simple product. Since we are making V small, let's call it $\Delta(V)$ to remind us that we are talking about just a small quantity.

Equation:

$$\oint_{\Delta(v)} \rho(v) \, dV \rightarrow \rho \Delta(v)$$

And thus, Gauss' Law becomes:

Equation:

$$\begin{aligned}\oint_{s,} D \, dS &= \varepsilon \oint_{s,} E \, dS \\ &= \rho \Delta(V)\end{aligned}$$

or

Equation:

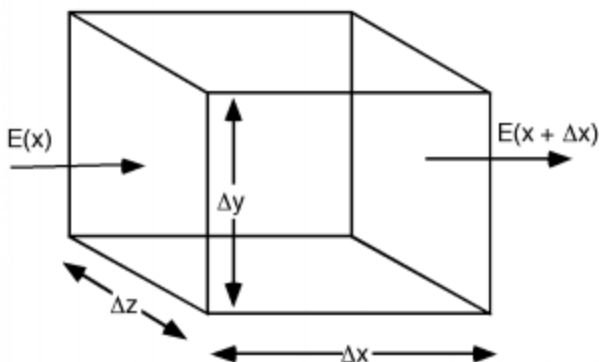
$$\frac{1}{\Delta(V)} \left(\oint_{s,} E \, dS \right) = \frac{\rho}{\varepsilon}$$

Now, by **definition** the limit of the LHS of [\[link\]](#) as $\Delta(V) \rightarrow 0$ is known as the divergence of the vector \mathbf{E} , $\text{div}(\mathbf{E})$. Thus we have

Equation:

$$\begin{aligned}\lim_{\Delta(V) \rightarrow 0} \frac{1}{\Delta(V)} \left(\oint_{s,} E \, dS \right) &= \text{div}(\mathbf{E}) \\ &= \frac{\rho}{\varepsilon}\end{aligned}$$

Note what this says about the divergence. The divergence of the vector \mathbf{E} is the limit of the surface integral of \mathbf{E} over a volume V , normalized by the volume itself, as the volume shrinks to zero. I like to think of as a kind of "point surface integral" of the vector \mathbf{E} .



Small volume for divergence

If \mathbf{E} only varies in one dimension, which is what we are working with right now, the expression for the divergence is particularly simple. It is easy to work out what it is from a simple picture. Looking at [\[link\]](#) we see that if \mathbf{E} is only pointed along one direction (let's say x) and is only a function of x , then the surface integral of \mathbf{E} over the volume $\Delta(V) = \Delta(x)\Delta(y)\Delta(z)$ is particularly easy to calculate.

Equation:

$$\oint_{s,} \mathbf{E} \, dS = \mathbf{E}(x + \Delta(x))\Delta(y)\Delta(z) - \mathbf{E}(x)\Delta(y)\Delta(z)$$

Where we remember that the surface integral is defined as being positive for an outward pointing vector and negative for one which points into the volume enclosed by the surface. Now we use the definition of the divergence

Equation:

$$\begin{aligned} \text{div}(\mathbf{E}) &= \lim_{\Delta(V) \rightarrow 0} \frac{1}{\Delta(V)} \left(\oint_{s,} \mathbf{E} \, dS \right) \\ &= \lim_{\Delta(V) \rightarrow 0} \frac{(\mathbf{E}(x + \Delta(x)) - \mathbf{E}(x))\Delta(y)\Delta(z)}{\Delta(x)\Delta(y)\Delta(z)} \\ &= \lim_{\Delta(V) \rightarrow 0} \frac{\mathbf{E}(x + \Delta(x)) - \mathbf{E}(x)}{\Delta(x)} \\ &= \frac{\partial \mathbf{E}(x)}{\partial x} \end{aligned}$$

So, we have for the differential form of Gauss' law:

Equation:

$$\frac{\partial \mathbf{E}(x)}{\partial x} = \frac{\rho(x)}{\varepsilon}$$

Thus, in our case, the rate of change of E with x , $\frac{d}{dx}(E)$, or the **slope of** $E(x)$ is just equal to the charge density, $\rho(x)$, divided by ε .

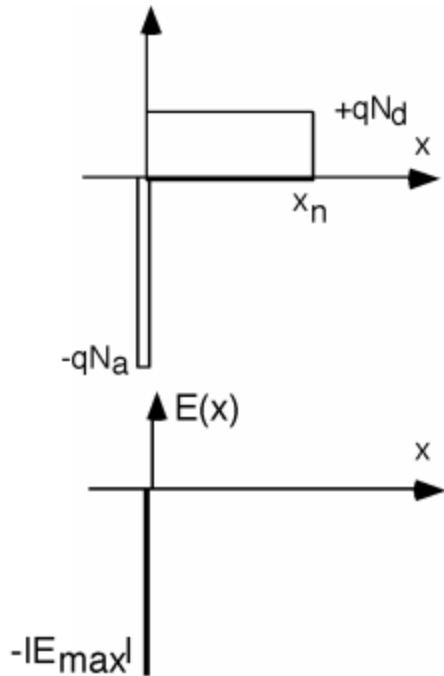
Depletion Width

We can now go back to the [charge density as a function of position graph](#) and easily find the electric field in the depletion region as a function of position. If we integrate [Gauss' Law](#), we get for the electric field:

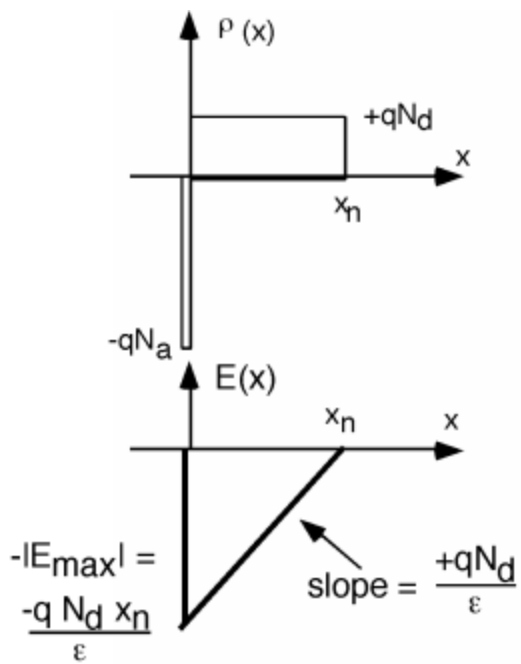
Equation:

$$\mathbf{E}(x) = \frac{1}{\epsilon} \int \rho(x) \, dx$$

We **could** write down an expression for $\rho(x)$ and then formally integrate it to get $\mathbf{E}(x)$ but we can also just do it graphically, which is a lot easier, and gives us a much more intuitive feeling for what is going on. Let's start doing our integral at [x equals -infinity] Whenever we perform an integral such as [\[link\]](#), we've got to remember to add a constant to our answer. Since we can not have an electric field which extends to infinity (either plus or minus) however, we can safely assume $\mathbf{E}(-\text{infinity}) = 0$ and remains at that value until we get to the edge of the depletion region at (essentially) x equals zero. Since the charge density is zero all the way up to the edge of depletion region, Gauss tells us that the electric field can not change here either. When we get to x=0 we encounter the large negative delta-function of negative charge at the edge of the depletion region. If you can remember back to your calculus, when you integrate a delta function, you get a step. Since the charge in the p-side delta function is negative, when we integrate it, we get a negative step. Since we don't know (yet) how big the step will be, let's just call it $-|E_{\text{max}}|$.



Finding the electric field in the
p-type region



Finishing the integral

In the n-side of the depletion region

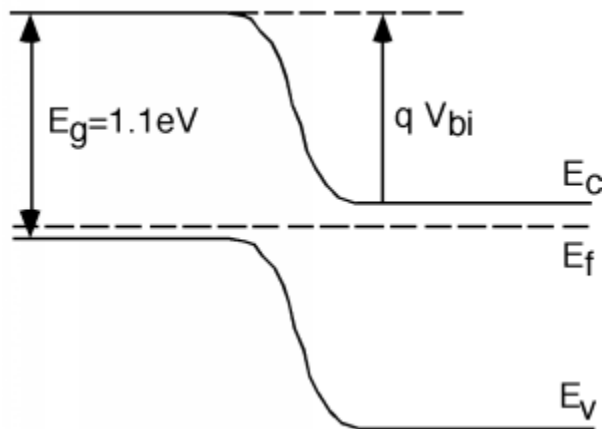
Equation:

$$\begin{aligned}\rho(x) &= (q)N_d \\ &= \varepsilon \frac{\partial E}{\partial x}\end{aligned}$$

and so we plot $E(x)$ with a (positive) slope of $\frac{qN_d}{\varepsilon}$, starting at $E(x) = -E_{\max}$ at $x = 0$. This line continues with this positive slope until it reaches a value of 0 at $x = x_n$. We know that $E(x)$ must equal 0 at $x = x_n$ because there is no further charge outside of the depletion region and E must be 0 outside this region.

We are now done doing the integral. We would know everything about this problem, if we just knew what x_n was. Note that since we know the slope of the triangle now, we can find $-E_{\max}$ in terms of the slope and x_n . We can derive an expression for x_n , if we remember that the integral of the electric field over a distance is the potential drop across that distance. What is the potential drop in going from the p-side to the n-side of the diode?

As a reminder, [\[link\]](#) shows the junction band diagram again. The potential drop must just be V_{bi} the "built-in" potential of the junction. Obviously V_{bi} can not be greater than 1.1 V, the band-gap potential. On the other hand, by looking at [\[link\]](#), and remembering that the bandgap in silicon is 1.1 eV, it will not be some value like 0.2 or 0.4 volts either. Let's make life easy for ourselves, and say $V_{bi} = 1$ Volt. This will not be too far off, and as you will see shortly, the answer is not very sensitive to the **exact** value of V_{bi} anyway.



Band diagram for a p-n junction

The integral of $\mathbf{E}(x)$ is now just the area of the triangle in [\[link\]](#). Getting the area is easy:

Equation:

$$\begin{aligned}
 \text{area} &= \frac{1}{2} \text{base} \times \text{height} \\
 &= \frac{1}{2} x_n \frac{q N_d x_n}{\epsilon} \\
 &= \frac{q N_d x_n^2}{2\epsilon} \\
 &= V_{bi}
 \end{aligned}$$

We can simply turn [\[link\]](#) around and solve for x_n .

Equation:

$$x_n = \sqrt{\frac{2\epsilon V_{bi}}{q N_d}}$$

As we said, for silicon, $\epsilon_{Si} = 1.1 \times 10^{-12}$. Let's let $N_d = 10^{16} \text{cm}^{-3}$ donors. As we already know from before, $q = 1.6 \times 10^{-19}$ Coulombs. This

makes $x_n = 3.7 \times 10^{-5}$ cm or 0.37 μm long. Not a very wide depletion region! How big is $|E_{\text{max}}|$? Plugging in

Equation:

$$E_{\text{max}} = \frac{qN_d x_n}{\epsilon}$$

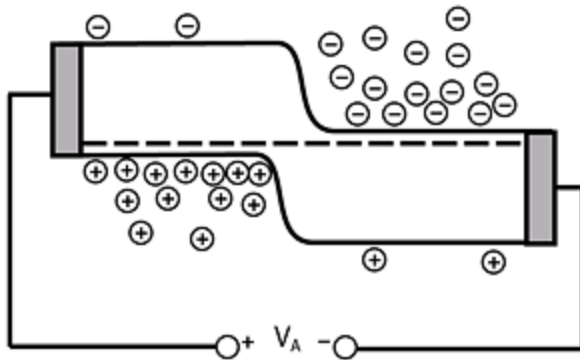
We find $|E_{\text{max}}| = 53,000 \frac{\text{V}}{\text{cm}}$! Why such a big electric field? Well, we've got to shift the potential by about a volt, and we do not have much distance to do it in (less than a micron), and so there must be, by default, a fairly large field in the depletion region. Remember, potential is electric field **times** distance.

Enough p-n junction electrostatics. The point of this exercise was two-fold; **a)** so you would know something about the details of what is really going on in a p-n junction **b)** to show you that with just some very simple electrostatics and a little thinking, it is not so hard to figure these things out!

Forward Biased PN Junctions

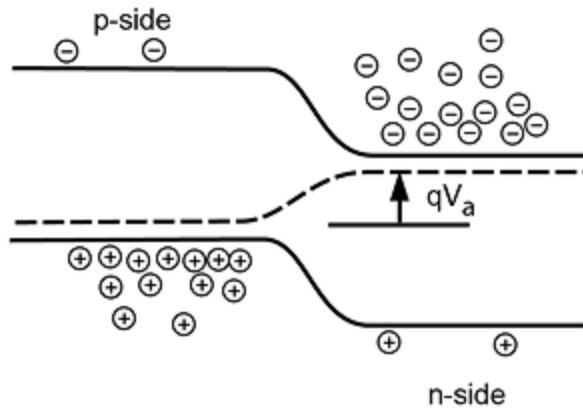
Now let's take a look at what happens when we apply an external voltage to this junction. First we need some conventions. We make connections to the device using **contacts**, which we show as cross-hatched blocks. These contacts allow the free passage of current into and out of the device.

Current usually flows through wires in the form of electrons, so it is easy to imagine electrons flowing into or out of the n-region. In the p-region, when electrons flow **out** of the device **into** the wire, holes will flow into the p-region (so as to maintain continuity of current through the contact.) When electrons flow into the p-region, they will recombine with holes, and so we have the net effect of holes flowing out of the p-region.



A p-n diode with contacts and external bias

With the convention that a **positive applied voltage** means that the terminal connected to the p-region is positive with respect to the terminal connected to the n-region. This is easy to remember; "p is positive, n is negative". Let us try to figure out what will happen when we apply a positive applied voltage V_a . If V_a is positive, then that means that the potential energy for electrons on the p-side must be **lower** than it was under the equilibrium condition. We reflect this on the band diagram by **lowering** the bands on the p-side from where they were originally. This is shown in [\[link\]](#).



A p-n junction under forward bias

As we can see from [\[link\]](#), when the p-region is lowered a couple of things happen. First of all, the Fermi level (the dotted line) is no longer a flat line, but rather it bends upward in going from the p-region to the n-region. The amount it bends (and hence the amount of shift of the bands) is just given by qV_a , where the energy scale we are using for the band diagram is in **electron-volts** which, as we said before, is a common measure of potential energy when we are talking about electronic materials. The other thing we can notice is that the electrons on the n-side and the holes on the p-side now "see" a lower potential energy barrier than they saw when no voltage was applied. In fact, it looks as if a lot of electrons now have sufficient energy such that they could move across from the n-region and flow into the p-region. Likewise, we would expect to see holes moving across from the p-region into the n-region.

This flow of carriers across the junction will result in a current flow across the junction. In order to see how this current will behave with applied voltage, we have to use a result from statistical thermodynamics concerning the distribution of electrons in the conduction band, and holes in the valence band. We saw from our "cups" analogy, that the electrons tend to fill in the lowest states first, with fewer and fewer of them as we go up in energy. For most situations, a very good description of just how the electrons are distributed in energy is given by a simple exponential decay. (This comes

about from a statistical analysis of electrons, which belong to a class of particles called **Fermions**. Fermions have the properties that they are: **a)** indistinguishable from one another **b)** obey the **Pauli Exclusion Principle** which says that two Fermions can not occupy the same exact **state** (energy and spin) **c)** remain at some fixed total number N .)

If $n(E)$ tells us how many electrons there are with an energy greater than some value E_c then $n(E)$ is given simply as:

Equation:

$$n(E) = N_d e^{-\frac{E-E_c}{kT}}$$

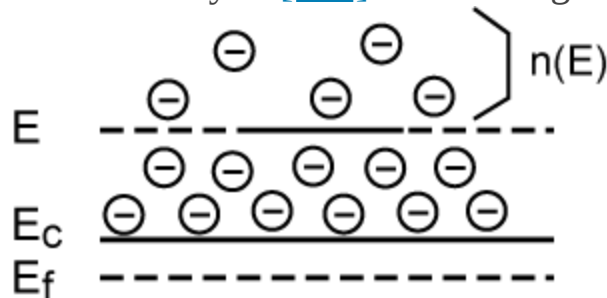
The expression in the denominator is just Boltzman's constant times the temperature in Kelvins. At room temperature kT has a value of about 1/40 of an eV or 25 meV. This number is sometimes called the **thermal voltage**, V_T , but it's ok for you to just think of it as a constant which comes from the thermodynamics of the problem. Because $kT \simeq 1/40$, you will sometimes see [\[link\]](#) and similar equations written as

Equation:

$$n(E) = N_d e^{-40(E-E_c)}$$

Which looks a little strange if you forget where the 40 came from, and just see it sitting there.

If the energy E is E_c the energy level of the conduction band, then $n(E_c) = N_d$, the density of electrons in the n-type material. As E increases above E_c , the density of electrons falls off exponentially, as depicted schematically in [\[link\]](#): Now let's go back to the unbiased junction.

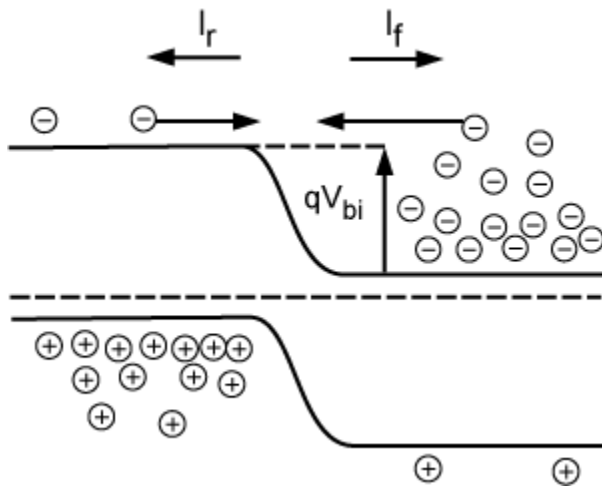


Distribution of electrons in the conduction band with energy

Remember, as we said before, there are currents flowing across the junction, even if there is no bias. The current we have shown as I_f is due to those electrons which have an energy greater than the built-in potential. They are flowing from right to left, as shown by the open arrow, which, of course, gives a current flowing from left to right, as shown by the solid arrows. Based on [\[link\]](#) the current should be proportional to:

Equation:

$$I_f \propto N_d e^{-\frac{qV_{bi}}{kT}}$$



Balanced flow across a junction

The principle of detailed balance says that at zero bias, $I_f = -I_r$ and so

Equation:

$$I_R \propto - \left(N_d e^{-\frac{qV_{bi}}{kT}} \right)$$

Equation:

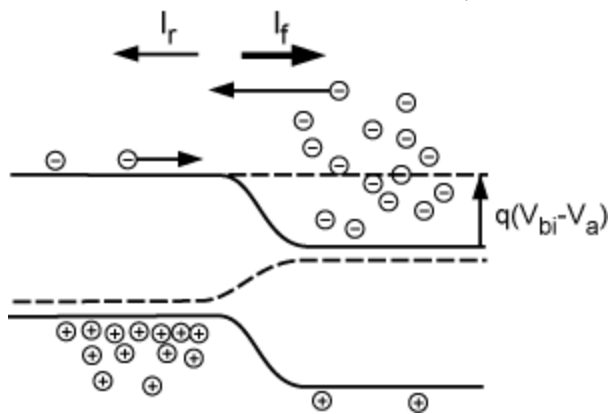
$$I_R = -(I_f \alpha) - N_d e^{-\frac{qV_{BI}}{kT}}$$

Now, what happens when we apply the bias? For the electrons over on the n-side, the barrier has been reduced from a height of qV_{bi} to $q(V_{bi} - V_a)$ and hence the forward current will be significantly increased.

Equation:

$$I_f \propto N_d e^{-\frac{q(V_{bi}-V_a)}{kT}}$$

The reverse current however, will remain just the [same as it was before](#).



Current when the junction is
forward biased

The total current across the junction is just $I_f + I_r$

Equation:

$$N_d \left(e^{\frac{qV_a}{kT}} - 1 \right)$$

where we have factored out the $N_d e^{-\frac{qV_{bi}}{kT}}$ term out of both expressions. We are not prepared, with what we know at this point, to get the other terms in the proportionality that are involved here. Also, the astute reader will note that we have not said anything about the holes, but it should be obvious that they will also contribute to the current, and the arguments we have made for electrons will hold for the holes just as well.

We can take the effect of the holes, and the other unknowns about the proportionality, and bind them all into one constant called I_{sat} , so that we write:

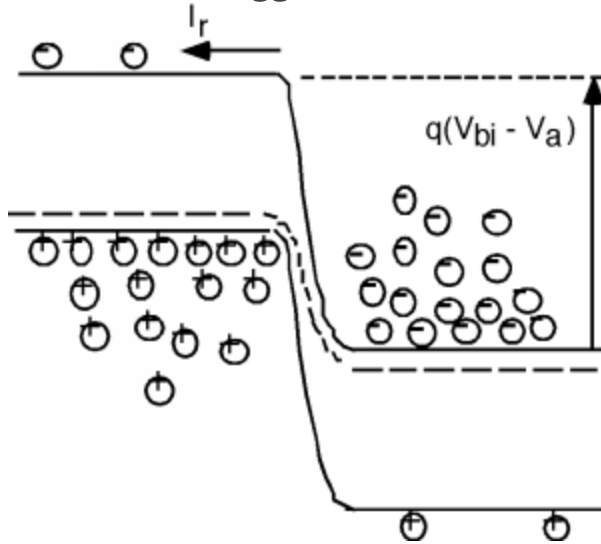
Equation:

$$I = I_{\text{sat}} \left(e^{\frac{qV_a}{kT}} - 1 \right)$$

This is the famous **diode equation** and is a very important result.

The Diode Equation

The reason for calling the proportionality constant I_{sat} will become obvious when we consider reverse bias. Let us now make V_a **negative** instead of positive. The applied electric field now **adds in the same direction** to the built-in field. This means the barrier will **increase** instead of decrease, and so we have what is shown in [\[link\]](#). Note that we have marked the barrier height as $q(V_{\text{bi}} - V_a)$ as before. It is just that now, V_a is negative, and so the barrier is bigger.



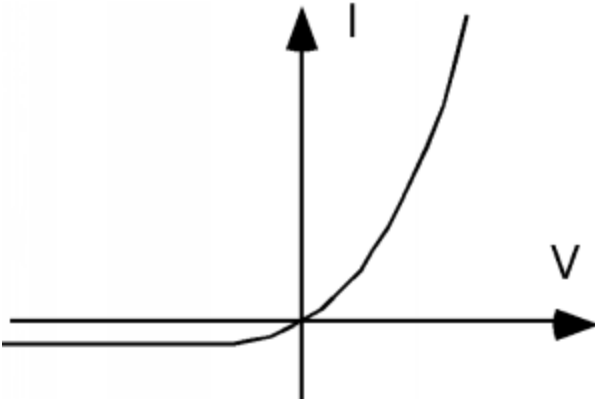
P-N junction under reverse bias
($V_a < 0$)

Remember, the electrons fall off exponentially as we move up in energy, so it does not take much of a shift of the bands before there are essentially **no** electrons on the n-side with enough energy to get over the barrier. This is reflected in the [diode equation](#) where, if we let V_a be a negative number, $e^{\frac{qV_a}{kT}}$ very quickly goes to zero and we are left with

Equation:

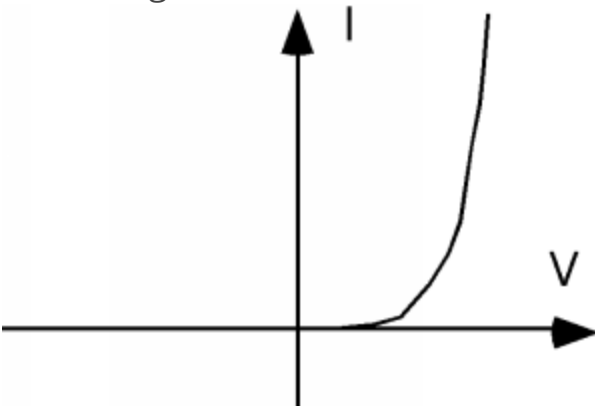
$$I = -I_{\text{sat}}$$

Thus, while in the forward bias direction, the current increases exponentially with voltage, in the reverse direction it simply saturates at $-I_{\text{sat}}$. A plot of I as a function of voltage or an **I-V characteristic curve** might look something like [\[link\]](#).



Idealized I-V curve for a p-n diode

In fact, for **real diodes** (ones made from silicon) I_{sat} is such a small value (on the order of 10^{-10} amps) that you can not even see it on most common measuring devices (oscilloscope, digital volt meter etc.) and if you were to look on a device called a **curve tracer** (which you will learn more about in Electronic Circuits [ELEC 342]) what you would really see would be something like [\[link\]](#).



Realistic I-V curve

We see what looks like zero current in the reverse direction, and in fact, what appears to be no current until we get a certain amount of voltage across the diode, after which it very quickly "turns on" with a very rapidly increasing forward current. For silicon, this "turn on" voltage is about 0.6 to 0.7 volts.

Digital volt meters (DVM's) use this characteristic for their "diode check" function. What they do is, when the "red" or positive lead is connected to the p-side (anode, or arrow in the diagram) and the "black" or negative lead is connected to the n-side (cathode, or bar in the diagram) of a diode, the meter attempts to pass (usually) 1 mA of current through the diode. If the 1 mA of current is allowed to flow, the meter then indicates the amount of forward voltage developed across the diode. If it reads something like 0.673 volts, then you can be pretty sure the diode is OK. Reverse the leads, and the diode is reverse biased, and the meter should read "OL" (overload) or something like that to indicate that no current is flowing.

[The diode equation](#) is usually approximated by two somewhat simpler equations, depending upon whether the diode is forward or reverse biased:
Equation:

$$I \simeq \begin{cases} 0 & \text{if } V_a < 0 \\ I_{\text{sat}} e^{\frac{qV_a}{kT}} & \text{if } V_a > 0 \end{cases}$$

For reverse bias, as we said, the current is essentially nil. In the forward bias case, the exponential term quickly gets much larger than unity, and so we can forget the "-1" term in the [diode equation](#). Remember, we said that kT at room temperature had a value of about 1/40 of an eV, so $\frac{q}{kT} \simeq 40V^{-1}$, this means we can also say for forward bias that

Equation:

$$I = I_{\text{sat}} e^{40V_a}$$

From this equation it is easy to see that only a small positive value for V_a is needed in order to make the exponential much greater than unity.

Now let's connect this "ideal diode equation" to the real world. One thing you might ask yourself is "How could I check to see if an actual diode follows the equation given [here](#)?" As we said, I_{sat} is a very small current, and so trying to do the reverse test is probably not going to be successful. What is usually done is to measure the diode current (and forward voltage) over several orders of magnitude of current.

Note: While the current can vary by many orders of magnitude, the voltage is more or less limited to values between 0 and 0.6 to 0.7 volts, not by any fundamental process, but rather simply by the fact that too much forward current will burn up the diode.

If we take the natural log of both sides of the second piece of [\[link\]](#), we find:

Equation:

$$\ln(I) = \ln(I_{\text{sat}}) + \frac{qV_a}{kT}$$

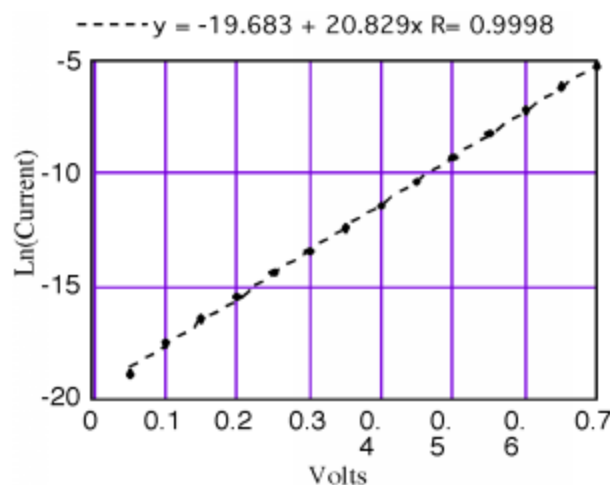
Thus, a plot of $\ln(I)$ as a function of V_a should yield a straight line with a slope of $\frac{q}{kT}$, or 40.

Well, I went into the lab, grabbed a real diode and made some measurements. [\[link\]](#) is a plot of the natural log of the current as a function of voltage from 0.05 to 0.70 volts. Included with this plot, is a linear curve fit to the data which is plotted as a dotted line. The linear fit goes through the data points quite nicely, so the current is surely an exponential function of the applied voltage! From the expression for the best fit, which is printed above the graph, we see that $\ln(I_{\text{sat}}) = -19.68$. That means that $I_{\text{sat}} = e^{-19.68} = 2.89 \times 10^{-9}$ amps, which is indeed a very small current.

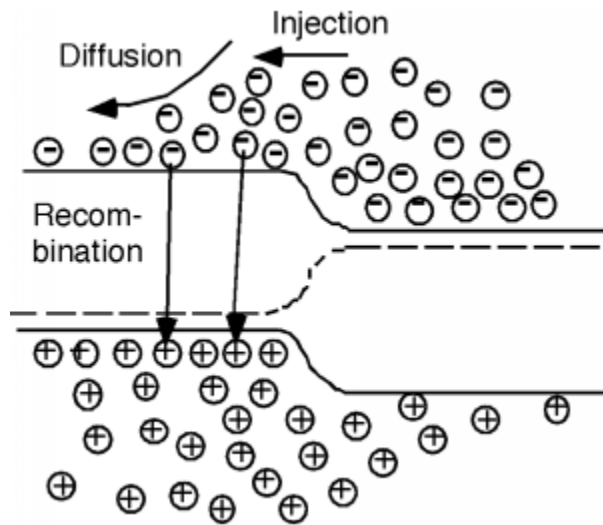
Look at the slope however. Its supposed to be 40, and yet it turns out to be slightly more than 20! This comes about because of some complex details of exactly what happens to the electrons and holes when they cross the junction. In what is called the **diffusion dominated situation** electrons and holes are injected across the junction, after which they diffuse away from the junction, and also recombine, until eventually they are all gone. This is shown schematically in [\[link\]](#). The other regime is called **recombination dominated** and here, the majority of the current is made up of the electrons and holes recombining directly with each other at the junction. This is shown in [\[link\]](#). For recombination dominated diode behavior, it turns out that the current is given by

Equation:

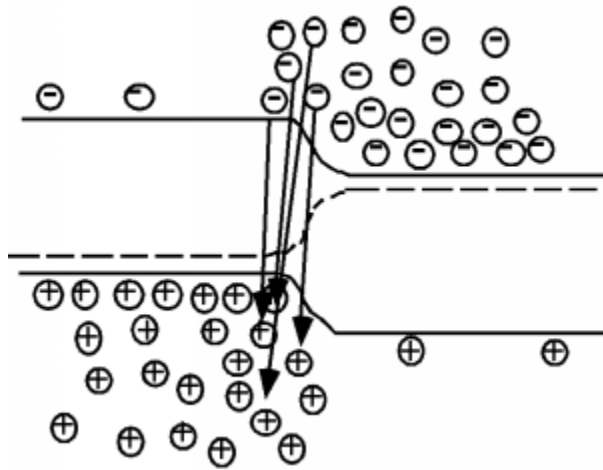
$$I = I_{\text{sat}} e^{\frac{qV_a}{2kT}}$$



Plot showing $\ln(I)$ as a function of V_a for a 1N4123 silicon diode



Diffusion dominated diode
behavior



Recombination dominated
diode behavior

In general, a particular diode might have a combination of these two effects going on, and so people often use a more general form for the diode equation:

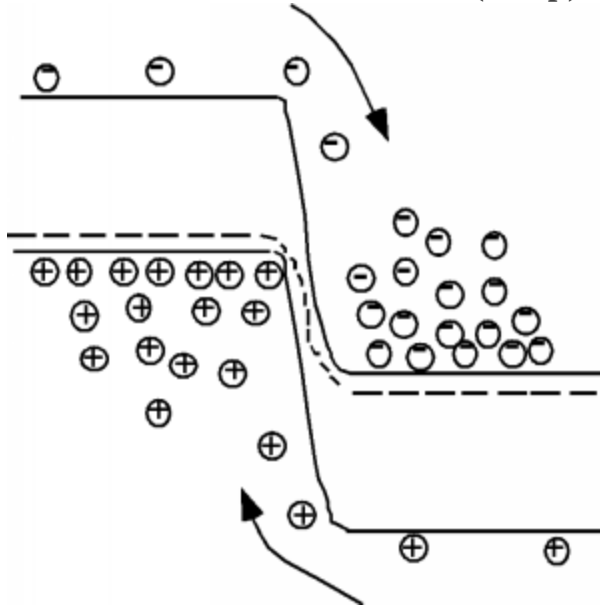
Equation:

$$I = I_{\text{sat}} e^{\frac{qV_a}{nkT}}$$

where n is called the **ideality factor** and is a number somewhere between 1 and 2. For the diode which gave the data for our example $n = 1.92$ and so most of the current is dominated by recombination of electrons and holes in the depletion region.

Reverse Biased/Breakdown

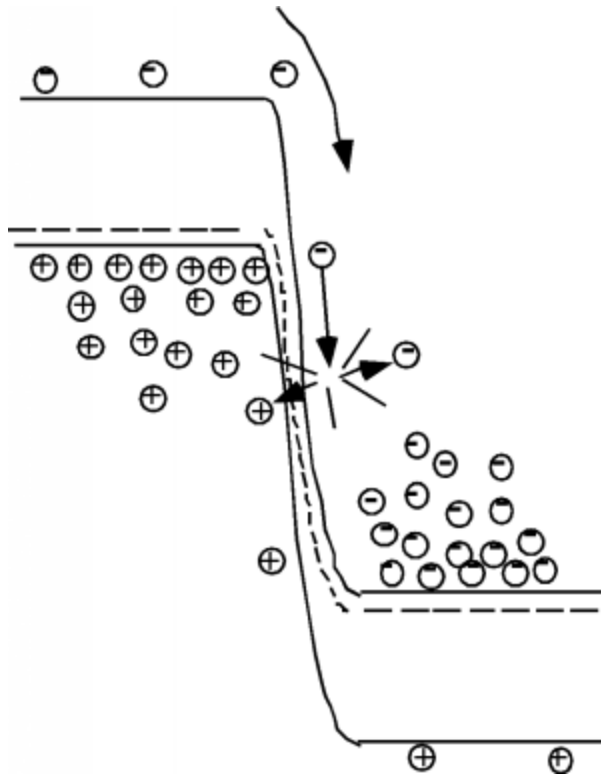
Before we leave diodes, it would be worthwhile exploring some other modes of operation, as well as some specific applications which will be of interest. We said that when the diode was reverse-biased (p-region negative with respect to the n-region) that the only current which flows is the reverse saturation current, resulting from the few thermally generated minority carriers which can fall down (or up) the barrier ([link](#)).



Reverse saturation current

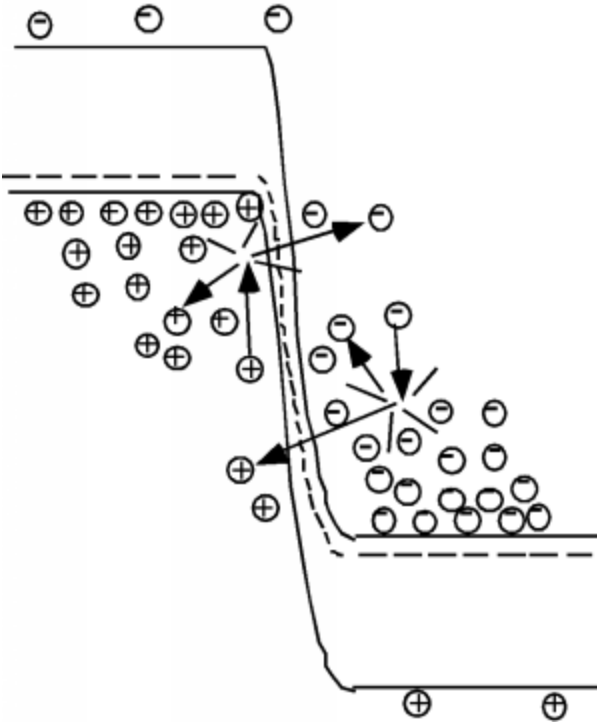
If we make the reverse bias even greater, the same current flows, but the carriers pick up more energy as they fall down the (now larger) junction potential. As they do this, it is possible for them to pick up so much energy, that when they collide with a lattice site, they create an additional electron-hole pair through a process called **impact ionization** ([link](#)). When this occurs, we now have current consisting of two electrons and one hole. These additional carriers can themselves collide and generate additional electron hole pairs as well. The current now consists of five electrons and two holes. This process is called **avalanche multiplication** ([link](#)), because we start with one carrier, and through a succession of impacts create more and more current. This process can in fact run away, much like an

avalanche on a snowy mountain side, in a process called **avalanche breakdown**.

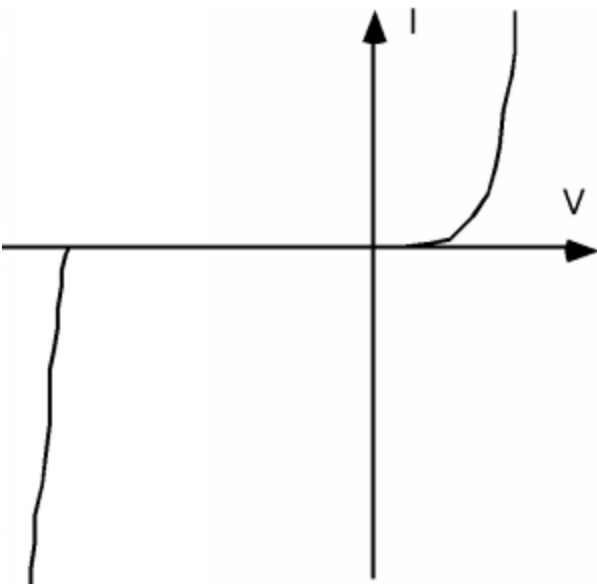


Impact Ionization

The net effect is to change the reverse characteristics of the diode somewhat. If we include the effect of breakdown in the I-V curve for the diode, we would see something like that in [\[link\]](#).

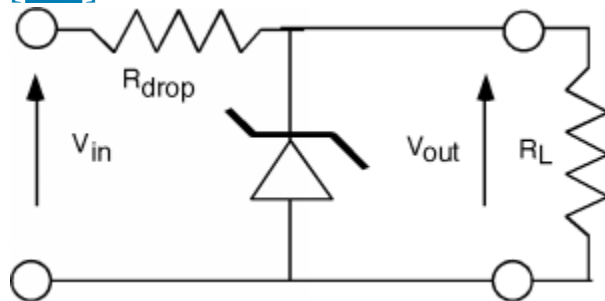


Avalanche multiplication



Diode I-V Curve showing both the forward characteristics and reverse breakdown

There is now a sudden onset of current after the avalanche breakdown voltage has been exceeded. Do not be confused into thinking that this "breakdown" means that the diode has been damaged. The process of avalanching itself is not destructive. But as you can see from [\[link\]](#), the diode current increases very rapidly once the breakdown threshold has been exceeded. Thus, if there is not something in series with the diode to limit the maximum current through it, it could be damaged by overheating. Diodes in breakdown are used as voltage references (the voltage across them is more or less independent of the current running through them) but you will always find a series current limiting resistor used along with them. Such diodes are called **Zener Diodes** (named after the grandfather of Will Rice's George Zener who graduated a few years ago...that is George did, not his grandfather) but the name is kind of a misnomer. The **Zener Effect** is also a reverse breakdown phenomena, but comes from direct field generation of extra carriers, rather than as a result of impact ionization. In truth, you can not tell the one effect from the other by looking at the diode I-V curve, and so all diodes used in reverse breakdown are called Zener Diodes. A circuit using a Zener diode as a voltage reference is shown in [\[link\]](#).



Voltage regulator circuit

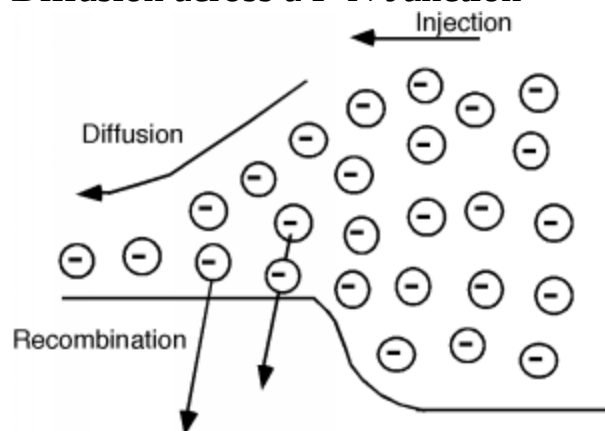
Diffusion

Introduction

Let us turn our attention to what happens to the electrons and holes, once they have been injected across a forward-biased junction. We will concentrate just on the electrons which are injected into the p-side of the junction, but keep in mind that similar things are also happening to the holes which enter the n-side.

As we saw a while back, when electrons are injected across a junction, they move away from the junction region by a diffusion process, while at the same time, some of them are disappearing because they are minority carriers (electrons in basically p-type material) and so there are lots of holes around for them to recombine with. This is all shown schematically in [\[link\]](#).

Diffusion across a P-N Junction

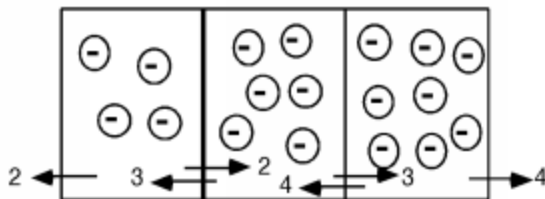


Processes involved in electron transport across a p-n junction

Diffusion Process Quantified

It is actually fairly easy to quantify this, and come up with an expression for the electron distribution within the p-region. First we have to look a little bit

First example of a diffusion problem



The diagram shows three adjacent rectangular boxes, each containing several circles with a minus sign inside, representing negative charges. The boxes are arranged horizontally. Below the boxes, arrows indicate the direction of the electric field at various points:

- Below the first box (left): An arrow points left, labeled '4'.
- Between the first and second boxes: Two arrows point left, labeled '6' and '4'.
- Below the second box (middle): An arrow points left, labeled '8'.
- Between the second and third boxes: Two arrows point right, labeled '6' and '8'.
- Below the third box (right): An arrow points right, labeled '8'.

Equation:

$$\text{Flux} = (-D_e) \frac{d n(x)}{d x}$$

Where D_e is simply a proportionality constant called the **diffusion coefficient**. Since we are talking about the motion of electrons, this diffusion flux must give rise to a current density $J_{e\text{diff}}$. Since an electron has a charge $-q$ associated with it,

Equation:

$$J_{e\text{diff}} = qD_e \frac{d n}{d x}$$

Now we have to invoke something called the **continuity equation**. Imagine we have a volume V which is filled with some charge Q . It is fairly obvious that if we add up all of the current density which is flowing out of the volume that it must be equal to the time rate of decrease of the charge within that volume. This idea is expressed in the formula below which uses a **closed-surface integral**, along with all the other integrals to follow:

Equation:

$$\oint_S J \, d S = - \frac{d Q}{d t}$$

We can write Q as

Equation:

$$Q = \int_V \rho(v) \, d V$$

where we are doing a volume integral of the charge density ρ over the volume V . Now we can use Gauss' theorem which says we can replace a surface integral of a quantity with a volume integral of its divergence:

Equation:

$$\oint_S J \, d S = \int \text{div} (J) \, d V$$

So, combining [\[link\]](#), [\[link\]](#) and [\[link\]](#), we have (note we are still dealing with surface and volume integrals):

Equation:

$$\int \operatorname{div} (J) \, dV = - \int \frac{d\rho}{dt} \, dV$$

Finally, we let the volume V shrink down to a point, which means the quantities inside the integral must be equal, and we have the differential form of the continuity equation (in one dimension)

Equation:

$$\begin{aligned} \operatorname{div} (J) &= \frac{\partial J}{\partial x} \\ &= - \frac{d\rho(x)}{dt} \end{aligned}$$

What about the Electrons?

Now let's go back to the electrons in the diode. The electrons which have been injected across the junction are called **excess minority carriers**, because they are electrons in a p-region (hence minority) but their concentration is greater than what they would be if they were in a sample of p-type material at equilibrium. We will designate them as n' , and since they could change with both time and position we shall write them as $n'(x, t)$. Now there are two ways in which $n'(x, t)$ can change with time. One would be if we were to stop injecting electrons in from the n-side of the junction. A reasonable way to account for the decay which would occur if we were not supplying electrons would be to write:

Equation:

$$\frac{d}{dt} n'(x, t) = - \frac{n'(x, t)}{\tau_r}$$

Where τ_r called the **minority carrier recombination lifetime**. It is pretty easy to show that if we start out with an excess minority carrier concentration n_o' at $t = 0$, then $n'(x, t)$ will goes as

Equation:

$$n'(x, t) = n'_0 e^{\frac{-t}{\tau_r}}$$

But, the electron concentration can also change because of electrons flowing into or out of the region x . The electron concentration $n'(x, t)$ is just $\frac{\rho(x, t)}{q}$. Thus, due to electron flow we have:

Equation:

$$\begin{aligned} \frac{d}{dt} n'(x, t) &= \frac{1}{q} \frac{d\rho(x, t)}{dt} \\ &= \frac{1}{q} \text{div} (J(x, t)) \end{aligned}$$

But, we can get an expression for $J(x, t)$ from [\[link\]](#). Reducing the divergence in [\[link\]](#) to one dimension (we just have a $\frac{\partial J}{\partial x}$) we finally end up with

Equation:

$$\frac{d}{dt} n'(x, t) = D_e \frac{d^2 n'(x, t)}{dx^2}$$

Combining [\[link\]](#) and [\[link\]](#) (electrons will, after all, suffer from both recombination and diffusion) and we end up with:

Equation:

$$\frac{d}{dt} n'(x, t) = D_e \frac{d^2 n'(x, t)}{dx^2} - \frac{n'(x, t)}{\tau_r}$$

This is a somewhat specialized form of an equation called the **ambipolar diffusion equation**. It seems kind of complicated but we can get some nice

results from it if we make some simply boundary condition assumptions. Let's see what we can do with this.

Using the Ambipolar Diffusion Equation

For anything we will be interested in, we will only look at **steady state solutions**. This means that the time derivative on the LHS of [\[link\]](#) is zero, and so we have (letting $n'(x, t)$ become simply $n'(x)$ since we no longer have any time variation to worry about)

Equation:

$$\frac{d^2}{dx^2} n'(x) - \frac{1}{D_e \tau_r} n'(x) = 0$$

Let's pick the not unreasonable boundary conditions that $n'(0) = n_0$ (the concentration of excess electrons just at the start of the diffusion region) and $n'(x) \rightarrow 0$ as $x \rightarrow \infty$ (the excess carriers go to zero when we get far from the junction) then

Equation:

$$n(x) = n_0 e^{-\frac{x}{\sqrt{D_e \tau_r}}}$$

The expression in the radical $\sqrt{D_e \tau_r}$ is called the **electron diffusion length**, L_e , and gives us some idea as to how far away from the junction the excess electrons will exist before they have more or less all recombined. This will be important for us when we move on to bipolar transistors.

Just so you can get a feel for some numbers, a typical value for the diffusion coefficient for electrons in silicon would be $D_e = 25 \frac{\text{cm}^2}{\text{sec}}$ and the minority carrier lifetime is usually around a microsecond. Thus

Equation:

$$\begin{aligned}
L_e &= \sqrt{D_e \tau_r} \\
&= \sqrt{25 \times 10^{-6}} \\
&= 5 \times 10^{-3} \text{ cm}
\end{aligned}$$

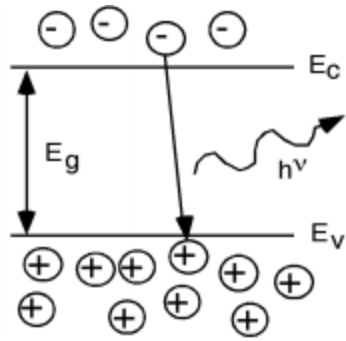
which is not very far at all!

Light Emitting Diode

Let's talk about the recombining electrons for a minute. When the electron falls down from the conduction band and fills in a hole in the valence band, there is an obvious loss of energy. The question is; where does that energy go? In silicon, the answer is not very interesting. Silicon is what is known as an **indirect band-gap material**. What this means is that as an electron goes from the bottom of the conduction band to the top of the valence band, it must also undergo a significant change in momentum. This all comes about from the details of the band structure for the material, which we will not concern ourselves with here. As we all know, whenever something changes state, we must still conserve not only energy, but also momentum. In the case of an electron going from the conduction band to the valence band in silicon, both of these things can only be conserved if the transition also creates a quantized set of lattice vibrations, called **phonons**, or "heat". Phonons possess **both** energy and momentum, and their creation upon the recombination of an electron and hole allows for complete conservation of both energy and momentum. All of the energy which the electron gives up in going from the conduction band to the valence band (1.1 eV) ends up in phonons, which is another way of saying that the electron heats up the crystal.

In some other semiconductors, something else occurs. In a class of materials called **direct band-gap semiconductors**, the transition from conduction band to valence band involves essentially no change in momentum. Photons, it turns out, possess a fair amount of energy (several eV/photon in some cases) but they have very little momentum associated with them. Thus, for a direct band gap material, the excess energy of the electron-hole recombination can either be taken away as heat, or more likely, as a photon of light. This **radiative transition** then conserves energy and momentum by giving off light whenever an electron and hole recombine. This gives rise to (for us) a new type of device, the light emitting diode (LED). Emission of a photon in an LED is shown schematically in [\[link\]](#).

Radiative recombination in a direct band-gap semiconductor



It was Planck who postulated that the energy of a photon was related to its frequency by a constant, which was later named after him. If the frequency of oscillation is given by the Greek letter "nu" (ν), then the energy of the photon is just $h\nu$, where h is Planck's constant, which has a value of 4.14×10^{-15} eV seconds.

Equation:

$$E = h\nu$$

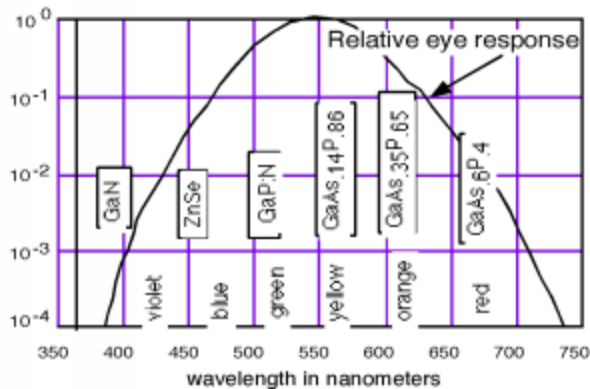
When we talk about light it is conventional to specify its wavelength, λ , instead of its frequency. Visible light has a wavelength on the order of nanometers (Red is about 600 nm, green about 500 nm and blue is in the 450 nm region.) A handy "rule of thumb" can be derived from the fact that $\lambda = \frac{c}{\nu}$, where c is the speed of light. Since $c = 3 \times 10^8 \frac{m}{sec}$ or $c = 3 \times 10^{17} \frac{nm}{sec}$

Equation:

$$\begin{aligned} \lambda(nm) &= \frac{hc}{E(eV)} \\ &= \frac{1242}{E(eV)} \end{aligned}$$

Thus, a semiconductor with a 2 eV band-gap should give off light at about 620 nm (in the red). A 3 eV band-gap material would emit at 414 nm, in the violet. The human eye, of course, is not equally responsive to all colors. We show this in [\[link\]](#), where we have also included the materials which are used for important light emitting diodes (LEDs) for each of the different spectral regions.

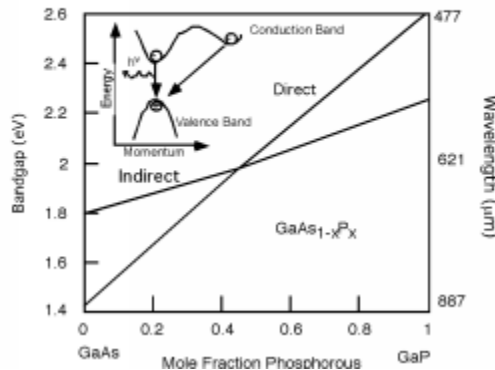
Relative response of the human eye to various colors



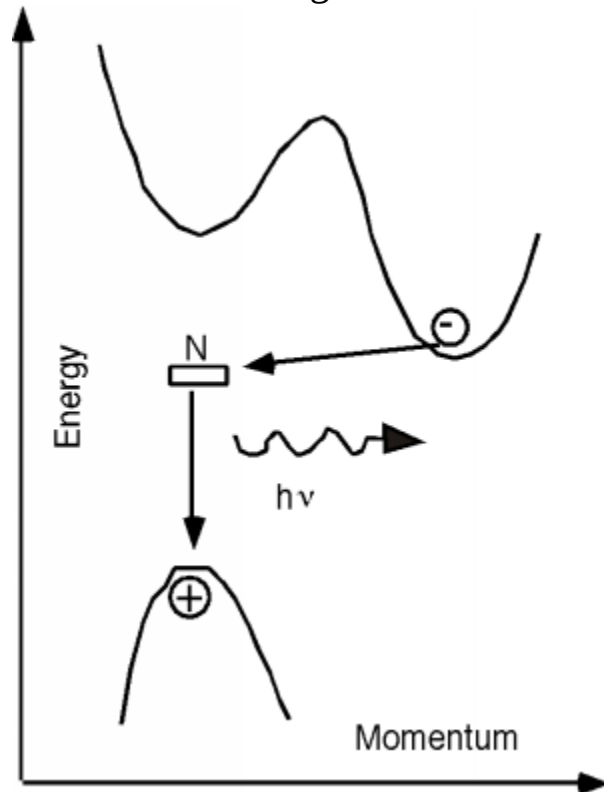
As you no doubt notice, a number of the important LEDs are based on the GaAsP system. GaAs is a direct band-gap semiconductor with a band gap of 1.42 eV (in the infrared). GaP is an indirect band-gap material with a band gap of 2.26 eV (550 nm, or green). Both As and P are group V elements. (Hence the nomenclature of the materials as **III-V compound semiconductors**.) We can replace some of the As with P in GaAs and make a mixed compound semiconductor $\text{GaAs}_{1-x}\text{P}_x$. When the mole fraction of phosphorous is less than about 0.45 the band gap is direct, and so we can "engineer" the desired color of LED that we want by simply growing a crystal with the proper phosphorus concentration! The properties of the GaAsP system are shown in [\[link\]](#). It turns out that for this system, there are actually **two** different band gaps, as shown in the [inset](#). One is a direct gap (no change in momentum) and the other is indirect. In GaAs, the direct gap has lower energy than the indirect one (like in the inset) and so the transition is a radiative one. As we start adding phosphorous to the system, both the direct and indirect band gaps increase in energy. However, the direct gap energy increases faster with phosphorous fraction than does the indirect one. At a mole fraction x of about 0.45, the gap energies cross over and the material goes from being a direct gap semiconductor to an indirect gap semiconductor. At $x = 0.35$ the band gap is about 1.97 eV (630 nm), and so we would only expect to get light up to the red using the GaAsP system for making LED's. Fortunately, people discovered that you could add an impurity (nitrogen) to the GaAsP system, which introduced a new level in the system. An electron could go from the indirect conduction band (for a mixture with a mole fraction greater than 0.45) to the nitrogen site, changing its momentum, but not its energy. It could then make a direct

transition to the valence band, and light with colors all the way to the green became possible. The use of a nitrogen **recombination center** is depicted in the [\[link\]](#).

Band gap for the GaAsP system



Addition of a nitrogen recombination center to indirect GaAsP



If we want colors with wavelengths shorter than the green, we must abandon the GaAsP system and look for more suitable materials. A compound semiconductor made from the II-VI elements Zn and Se make up one promising system, and several research groups have successfully made

blue and blue-green LEDs from ZnSe. SiC is another (weak) blue emitter which is commercially available on the market. Recently, workers at a tiny, unknown chemical company stunned the "display world" by announcing that they had successfully fabricated a blue LED using the II-V material GaN. A good blue LED has been the "holy grail" of the display and CD ROM research community for a number of years now. Obviously, adding blue to the already working green and red LED's completes the set of 3 primary colors necessary for a full-color flat panel display (Hang a TV screen on your wall like a picture?). Using a blue LED or laser in a CD ROM would more than quadruple its data capacity, as bit diameter scales as λ , and hence the area as λ^2 .

LASER

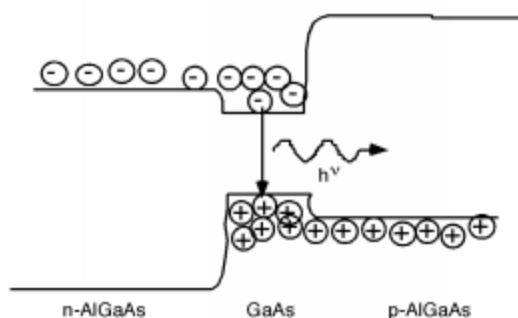
Speaking of lasers, what is the difference between an LED and a solid state laser? There are some differences, but both devices operate on the same principle of having excess electrons in the conduction band of a semiconductor, and arranging it so that the electrons recombine with holes in a radiative fashion, giving off light in the process. What is different about a laser? In an LED, the electrons recombine in a random and unorganized manner. They give off light by what is known as **spontaneous emission**, which simply means that the exact time and place where a photon comes out of the device is up to each individual electron, and things happen in a random way.

There is another way in which an excited electron can emit a photon however. If a field of light (or a set of photons) happens to be passing by an electron in a high energy state, that light field can induce the electron to emit an additional photon through a process called **stimulated emission**. The photon field **stimulates** the electron to emit its energy as an additional photon, which comes out **in phase with the stimulating field**. This is the big difference between **incoherent light** (what comes from an LED or a flashlight) and **coherent light** which comes from a laser. With coherent light, all of the electric fields associated with each photon are all exactly in phase. This coherence is what enables us to keep a laser beam in tight focus, and to allow it to travel a large distance without much divergence or spreading out.

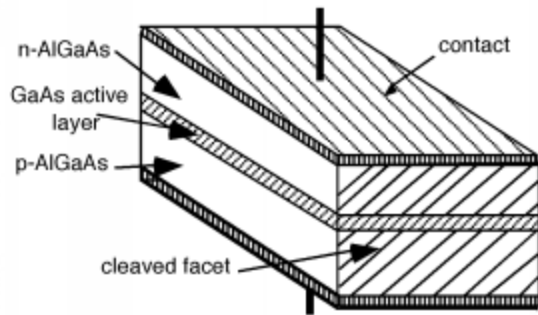
So how do we restructure an LED so that the light is generated by stimulated emission rather than spontaneous emission? Firstly, we build what is called a **heterostructure**. All this means is that we build up a sandwich of somewhat different materials, with different characteristics. In this case, we put two wide band-gap regions around a region with a narrower band gap. The most important system where this is done is the AlGaAs/GaAs system. A band diagram for such a set up is shown in [\[link\]](#). AlGaAs (pronounced "Al-Gas") has a larger band-gap than does GaAs. The potential "well" formed by the GaAs means that the electrons and holes will be confined there, and all of the recombination will occur in a very narrow strip. This greatly increases the chances that the carriers can interact, but we

still need some way for the photons to behave in the proper manner. [\[link\]](#) is a picture of what a real diode might look like. We have the active GaAs layer sandwich in-between the two heterostructure confinement layers, with a contact on top and bottom. On either end of the device, the crystal has been "cleaved" or broken along a crystal lattice plane. This results in a shiny "mirror-like" surface, which will reflect photons. The back surface (which we can not see here) is also cleaved to make a mirror surface. The other surfaces are purposely roughened so that they do not reflect light. Now let us look at the device from the side, and draw just the band diagram for the GaAs region ([\[link\]](#)). We start things off with an electron and hole recombining spontaneously. This emits a photon which heads towards one of the mirrors. As the photon goes by other electrons, however, it may cause one of them to decay by stimulated emission. The two (in phase) photons hit the mirror and are reflected and start back the other way . As they pass additional electrons, they stimulate them into a transition as well, and the optical field within the laser starts to build up. After a bit, the photons get down to the other end of the cavity. The cleaved facet, while it acts like a mirror, is not a perfect one. Some light is not reflected, but rather "leaks"; though, and so becomes the output beam from the laser. The details of finding what the ratio of reflected to transmitted light is will have to wait until later in the course when we talk about dielectric interfaces. The rest of the photons are reflected back into the cavity and continue to stimulate emission from the electrons which continue to enter the gain region because of the forward bias on the diode.

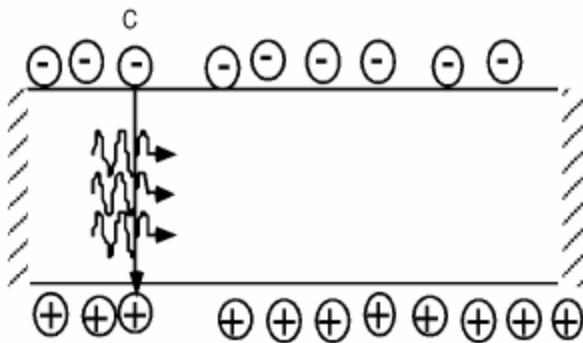
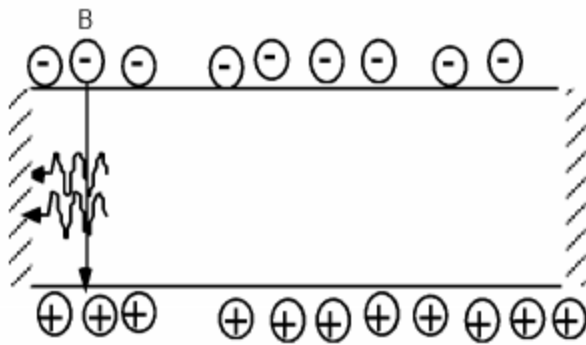
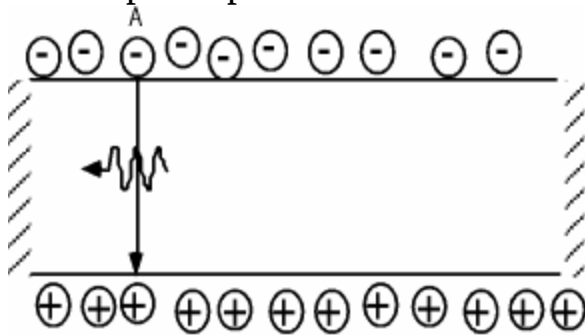
Double Heterostructure GaAs/AlGaAs laser



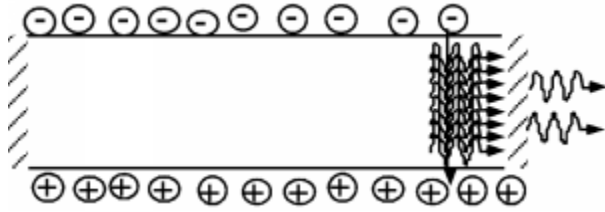
Laser Diode



Build up of a photon field in a laser diode



Output Coupling

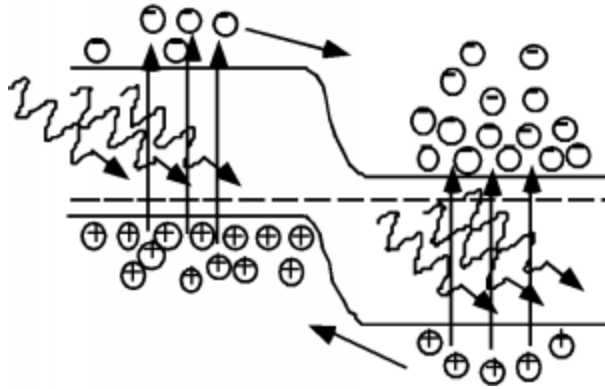


In reality, the photons do not move back and forth in a big "clump" as we have described here, rather they are distributed uniformly along the gain region. The field within the cavity will build up to the point where the loss of energy by light leaking out of the mirrors just equals the rate at which energy is replaced by the recombining electrons.

Solar Cells

Now let us look at the opposite process of light generation for a moment. Consider the following situation.

P-N diode under illumination



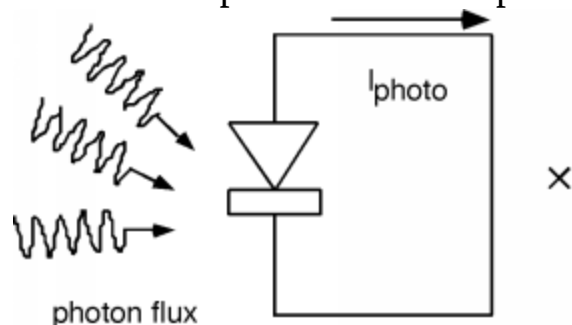
We have just a plain old normal p-n junction, only now, instead of applying an external voltage, we imagine that the junction is being illuminated with light whose photon energy is greater than the band-gap. In this situation, instead of recombination, we will get photo-generation of electron hole pairs. The photons simply excite electrons from the full states in the valence band, and "kick" them up into the conduction band, leaving a hole behind[\[footnote\]](#). As you can see from [\[link\]](#), this creates excess electrons in the conduction band in the p-side of the diode, and excess holes in the valence band of the n-side. These carriers can diffuse over to the junction, where they will be swept across by the built-in electric field in the depletion region. If we were to connect the two sides of the diode together with a wire, a current would flow through that wire as a result of the electrons and holes which move across the junction.

This is similar to the thermal excitation process we talked about earlier.

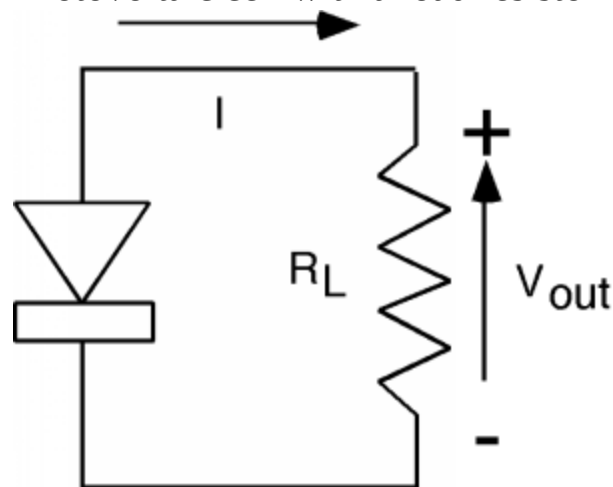
Which way would the current flow? A quick look at [\[link\]](#) shows that holes (positive charge carriers) are generated on the n-side and they float up to the p-side as they go across the junction. Hence positive current must be coming out of the anode, or p-side of the junction. Likewise, electrons generated on the p-side fall down the junction potential, and come out the n-side, but since they have negative charge, this flow represents current going **into** the cathode. We have constructed a **photovoltaic diode**, or **solar cell**! [\[link\]](#) is a picture of what this would look like schematically. We might

like to consider the possibility of using this device as a source of energy, but the way we have things set up now, since the voltage across the diode is zero, and since power equals current times voltage, we see that we are getting nada from the cell. What we need, obviously, is a load resistor, so let's put one in. It should be clear from [\[link\]](#) that the photo current flowing through the load resistor will develop a voltage which it biases the diode in the **forward** direction, which, of course will cause current to flow back into the anode. This complicates things, it seems we have current coming **out** of the diode and current going **into** the diode all at the same time! How are we going to figure out what is going on?

Schematic representation of a photovoltaic cell



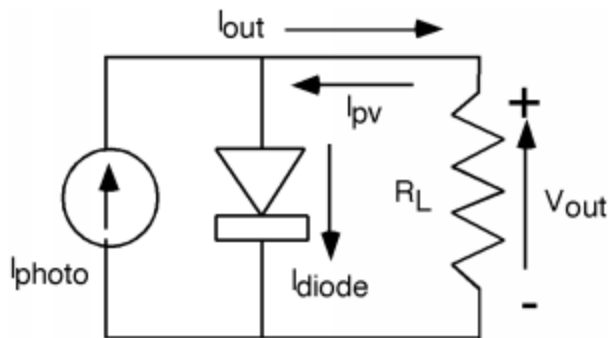
Photovoltaic cell with a load resistor



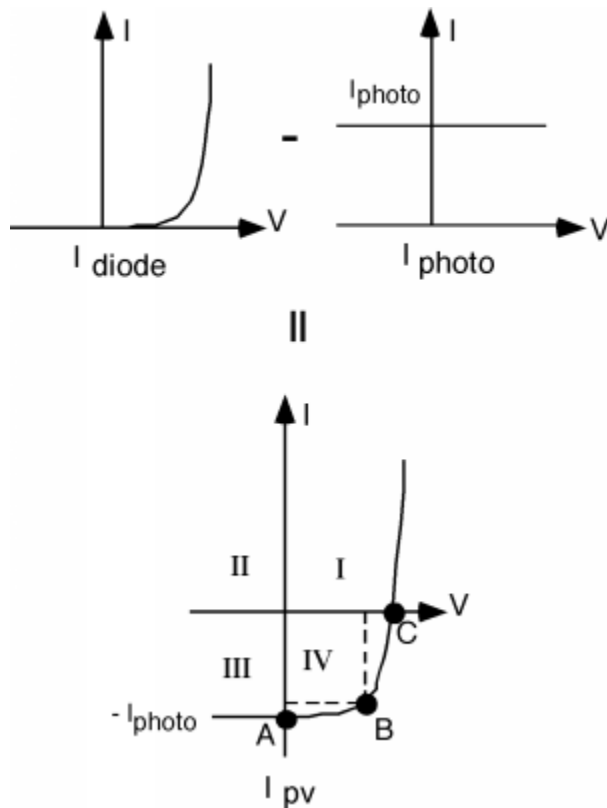
The answer is to make a model. The current which arises due to the photon flux can be conveniently represented as a current source. We can leave the diode as a diode, and we have the circuit shown in [\[link\]](#). Even though we

show I_{out} coming out of the device, we know by the usual polarity convention that when we define V_{out} as being positive at the top, then we should show the current for the photovoltaic, I_{pv} as current going into the top, which is what was done in [\[link\]](#). Note that $I_{\text{pv}} = I_{\text{diode}} - I_{\text{photo}}$, so all we need to do is to subtract the two currents; we do this graphically in [\[link\]](#). Note that we have numbered the four quadrants in the I-V plot of the total PV current. In quadrant I and III, the product of I and V is a positive number, meaning that power is being **dissipated** in the cell. For quadrant II and IV, the product of I and V is negative, and so we are getting power **from** the device. Clearly we want to operate in quadrant IV. In fact, without the addition of an external battery or current source, the circuit, will **only** run in the IV'th quadrant. Consider adjusting R_L , the load resistor from 0 (a short) to ∞ (an open). With $R_L = 0$, we would be at point A on [\[link\]](#). As R_L starts to increase from zero, the voltage across both the diode and the resistor will start to increase also, and we will move to point B, say. As R_L gets bigger and bigger, we keep moving along the curve until, at point C, where R_L is an open and we have the maximum voltage across the device, but, of course, no current coming out!

Model of PV cell



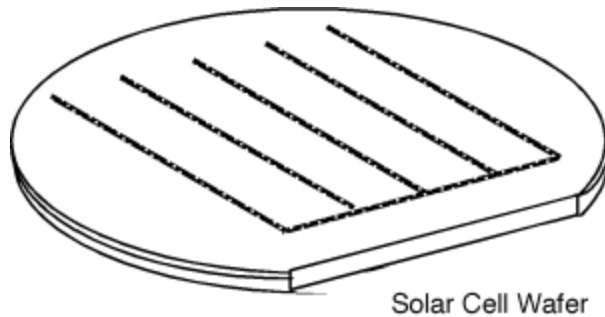
Combining the diode and the current source



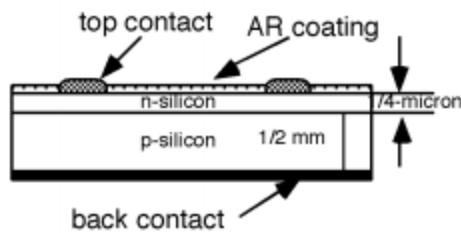
Power is VI so at B for instance, the power coming out would be represented by the area enclosed by the two dotted lines and the coordinate axes. Somewhere about where I have point B would be where we would be getting the most power out of our solar cell.

[\[link\]](#) shows you what a real solar cell would look like. They are usually made from a complete wafer of silicon, to maximize the usable area. A shallow ($0.25 \mu\text{m}$) junction is made on the top, and top contacts are applied as stripes of metal conductor as shown. An anti-reflection (AR) coating is applied on top of that, which accounts for the bluish color which a typical solar cell has.

A real solar cell



Solar Cell Wafer

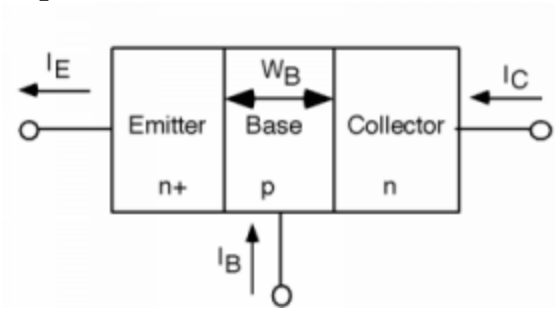


Side View

The solar power flux on the earth's surface is (conveniently) about $1 \frac{\text{kW}}{\text{m}^2}$ or $100 \frac{\text{mW}}{\text{cm}^2}$. So if we made a solar cell from a 4 inch diameter wafer (typical) it would have an area of about 81cm^2 and so would be receiving a flux of about 8.1 Watts. Typical cell efficiencies run from about 10% to maybe 15% unless special (and costly) tricks are made. This means that we will get about 1.2 Watts out from a single wafer. Looking at B on 2.59 we could guess that V_{out} will be about 0.5 to 0.6 volts, thus we could expect to get maybe around 2.5 amps from a 4 inch wafer at 0.5 volts with 15% efficiency under the illumination of one sun.

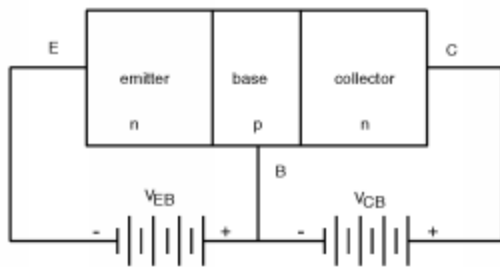
Introduction to Bipolar Transistors

Let's leave the world of two terminal devices (which are all called diodes by the way; diode just means two-terminals) and venture into the much more interesting world of three terminals. The first device we will look at is called the **bipolar transistor**. Consider the structure shown in [\[link\]](#):
Bipolar Transistor Structure

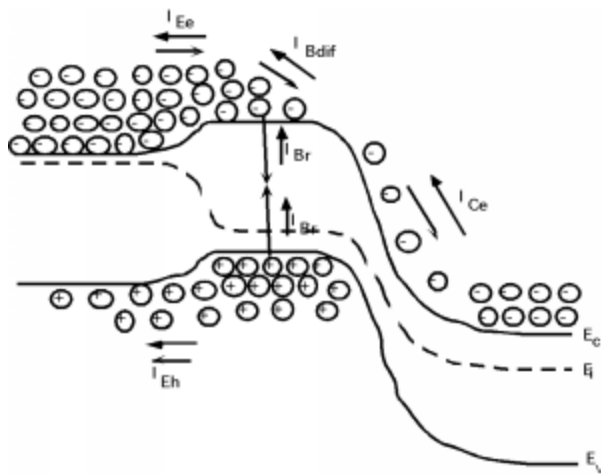


Structure of an NPN bipolar transistor

The device consists of three layers of silicon, a heavily doped n-type layer called the emitter, a moderately doped p-type layer called the base, and third, more lightly doped layer called the collector. In a biasing (applied DC potential) configuration called **forward active biasing**, the emitter-base junction is forward biased, and the base-collector junction is reverse biased. [\[link\]](#) shows the biasing conventions we will use. Both bias voltages are referenced to the base terminal. Since the base-emitter junction is forward biased, and since the base is made of p-type material, V_{EB} must be negative. On the other hand, in order to reverse bias the base-collector junction V_{CB} will be a positive voltage.
forward_active_biasing



Forward active biasing of an npn bipolar transistor



Band diagram and carrier fluxes in a bipolar transistor

Now, let's draw the band-diagram for this device. At first this might seem hard to do, but we know what forward and reverse biased band diagrams look like, so we'll just stick one of each together. We show this in [\[link\]](#). [\[link\]](#) is a very busy figure, but it is also very important, because it shows all of the important features in the operation the transistor. Since the base-emitter junction is forward biased, electrons will go from the (n-type) emitter into the base. Likewise, some holes from the base will be injected into the emitter.

In [\[link\]](#), we have two different kinds of arrows. The open arrows which are attached to the carriers, show us which way the carrier is moving. The solid arrows which are labeled with some kind of subscripted I , represent current flow. We need to do this because for holes, motion and current flow are in the same direction, while for electrons, carrier motion and current flow are in opposite directions.

Just as we saw in the last chapter, the electrons which are injected into the base diffuse away from the emitter-base junction towards the (reverse biased) base-collector junction. As they move through the base, some of the electrons encounter holes and recombine with them. Those electrons which **do** get to the base-collector junction run into a large electric field which sweeps them out of the base and into the collector. (They "fall" down the large potential drop at the junction.)

These effects are all seen in [\[link\]](#), with arrows representing the various currents which are associated with each of the carriers fluxes. I_{Ee} represents the current associated with the electron injection into the base. (It points in the opposite direction from the motion of the electrons, since electrons have a negative charge.) I_{Eh} represents the current associated with holes injection into the emitter from the base. I_{Br} represents recombination current in the base, while I_{Ce} represents the electron current going into the collector. It should be easy for you to see that:

Equation:

$$I_E = I_{Ee} + I_{Eh}$$

Equation:

$$I_B = I_{Eh} + I_{Br}$$

Equation:

$$I_C = I_{Ce}$$

In a "good" transistor, almost all of the current across the base-emitter junction consists of electrons being injected into the base. The transistor engineer works hard to design the device so that very little emitter current is

made up of holes coming from the base into the emitter. The transistor is also designed so that almost all of those electrons which are injected into the base make it across to the base-collector reverse-biased junction. Some recombination is unavoidable, but things are arranged so as to minimize this effect.

Transistor Equations

There are several "figures of merit" for the operation of the transistor. The first of these is called the **emitter injection efficiency**, γ . The emitter injection efficiency is just the ratio of the electron current flowing in the emitter to the total current across the emitter base junction:

Equation:

$$\gamma = \frac{I_e}{I_{Ee} + I_{Eh}}$$

If you go back and look at [the diode equation](#) you will note that the electron forward current across a junction is proportional to N_d the doping on the n-side of the junction. Clearly the hole current will be proportional to N_a , the acceptor doping on the p-side of the junction. Thus, at least to first order

Equation:

$$\gamma = \frac{N_{dE}}{N_{dE} + N_{aB}}$$

(There are some other considerations which we are ignoring in obtaining this expression, but to first order, and for most "real" transistors, [\[link\]](#) is a very good approximation.)

The second "figure of merit" is the base transport factor, α_T . The base transport factor tells us what fraction of the electron current which is injected into the base actually makes it to collector junction. This turns out to be given, to a very good approximation, by the expression

Equation:

$$\alpha_T = 1 - \frac{1}{2} \left(\frac{W_B}{L_e} \right)^2$$

Where W_B is the physical width of the base region, and L_e is the electron diffusion length, defined in the [electron diffusion length equation](#).

Equation:

$$L_e = \sqrt{D_e \tau_r}$$

Clearly, if the base is very narrow compared to the diffusion length, and since the electron concentration is falling off like $e^{\frac{-x}{L_e}}$ the shorter the base is compared to L_e the greater the fraction of electrons who will actually make it across. We saw before that a typical value for L_e might be on the order of 0.005 cm or 50 μm . In a typical bipolar transistor, the base width, W_B is usually only a few μm and so α can be quite close to unity as well.

Looking back at [this figure](#), it should be clear that, so long as the collector-base junction remains reverse-biased, the collector current I_{C_e} , will only depend on how much of the total emitter current actually gets collected by the reverse-biased base-collector junction. That is, the collector current I_C is just some fraction of the total emitter current I_E . We introduce yet one more constant which reflects the ratio between these two currents, and call it simply " α ." Thus we say

Equation:

$$I_C = \alpha I_E$$

Since the **electron** current into the base is just γI_E and α_T of that current reaches the collector, we can write:

Equation:

$$\begin{aligned} I_C &= \alpha I_E \\ &= \alpha_T \gamma I_E \end{aligned}$$

Looking back at the [structure of an npn bipolar transistor](#), we can use Kirchoff's current law for the transistor and say:

Equation:

$$I_C + I_B = I_E$$

or

Equation:

$$\begin{aligned} I_B &= I_E - I_C \\ &= \frac{I_C}{\alpha} - I_C \end{aligned}$$

This can be re-written to express I_C in terms of I_B as:

Equation:

$$I_C = \frac{\alpha}{1 - \alpha} I_B \equiv \beta I_B$$

This is the fundamental operational equation for the bipolar equation. It says that the collector current is dependent only on the base current. Note that if α is a number close to (but still slightly less than) unity, then β which is just given by

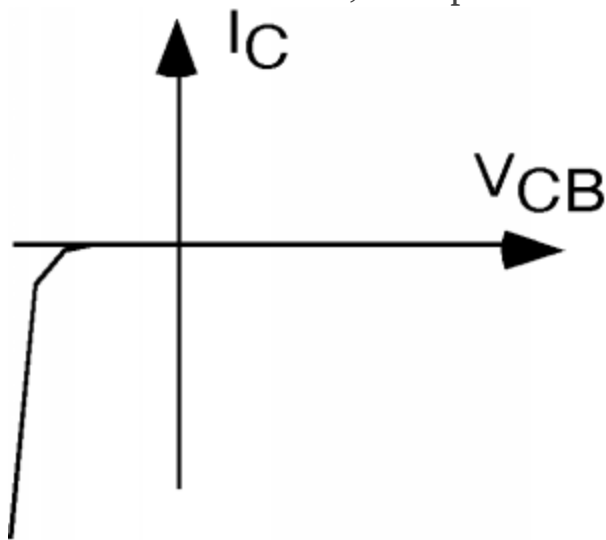
Equation:

$$\beta = \frac{\alpha}{1 - \alpha}$$

will be a fairly large number. Typical values for α will be on the order of 0.99 or greater, which puts β , the current gain, at around 100 or more! This means that we can control, or amplify the current going into the collector of the transistor with a current 100 times smaller going into the base. This all occurs because the ratio of the collector current to the base current is fixed by the conditions across the emitter-base junction, and the ratio of the two, I_C to I_B is always the same.

Transistor I-V Characteristics

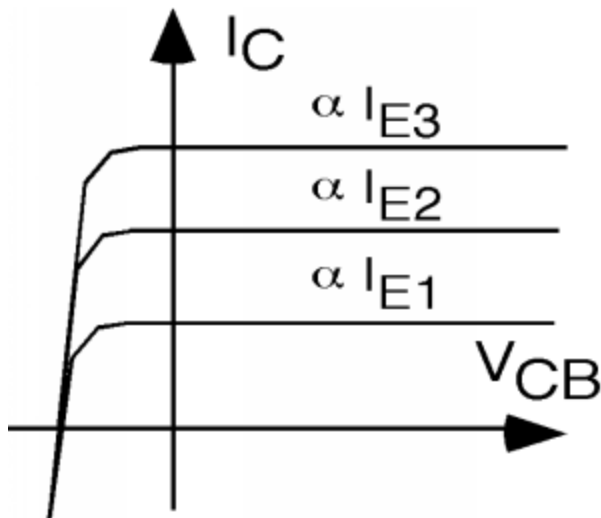
Let's now take a look at some current voltage relationships for the bipolar transistor. In the absence of any voltage or current on the emitter-base junction, if we were to make a plot of I_C as a function of V_{CB} it would look something like [\[link\]](#). Check back with the voltage convention in the figures on [the structure](#) and [forward active biasing](#) of a bipolar transistor to make sure you agree with what I drew. All we've got here is a pn junction or diode. It just happens to be biased in a reverse direction, so it conducts when V_{CB} is negative and not when V_{CB} is positive. Thus, all we need to do is draw a diode curve, but upside down!



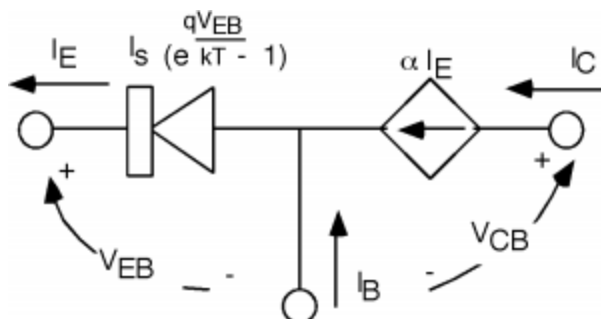
I-V for the collector-base
terminals of the bipolar
transistor

What happens if we now **also** have some bias applied to the emitter-base junction? As we saw, so long as the base-collector junction is reverse biased, almost all of the collector current consists of electrons which have been injected into the base by the emitter, diffuse across the base region, and then fall down the base-collector junction. The rate at which electrons fall down the junction does not depend on how large a drop there is (e.g. how big V_{CB} is). The only thing that matters, in so far as the collector

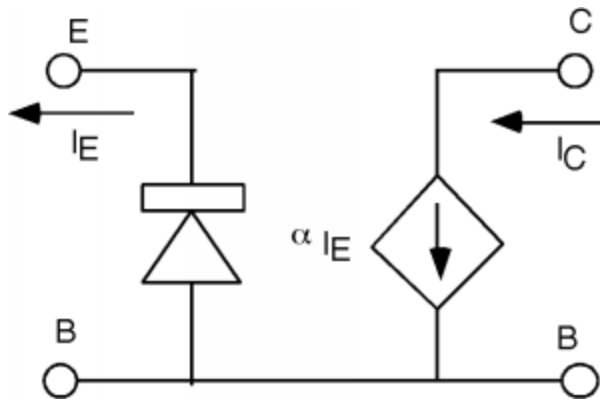
current is concerned, is how fast electrons are being injected into the base region, which is, of course, determined by the emitter current I_E . Thus for several different values of emitter current, I_{E1} , I_{E2} , and I_{E3} , we might see something like [\[link\]](#). In the first quadrant, which is in the "forward active bias mode," the output from the collector terminal looks more or less like a current source; that is I_C is a constant, regardless of what V_{CB} is. Note however, that we must use a **controlled source**, in this case, a current-controlled current source, since I_C depends on what I_E happens to be. Obviously, looking in the (forward biased) emitter-base terminal, we see the usual p-n junction. Thus, if we were interested in building a "model" of this device, we might come up with something like [\[link\]](#). Note that the base terminal is common to both inputs. Since we would actually like to think of the transistor as a two-port device (with an input and an output) the model for the transistor is often drawn as shown in [\[link\]](#).



Common base characteristics of the bipolar transistor

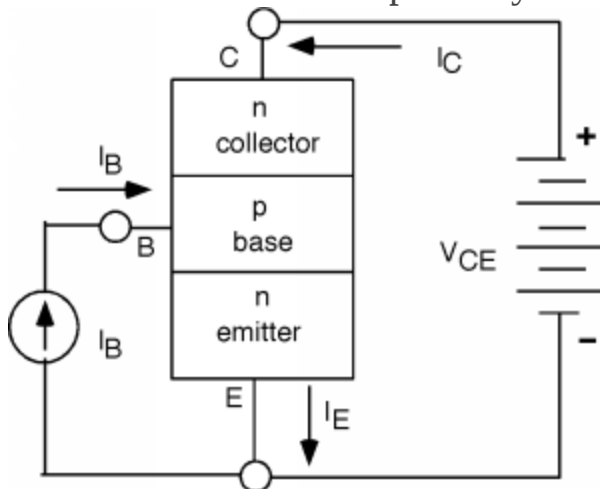


Model for the common base transistor



Re-drawn common base transistor

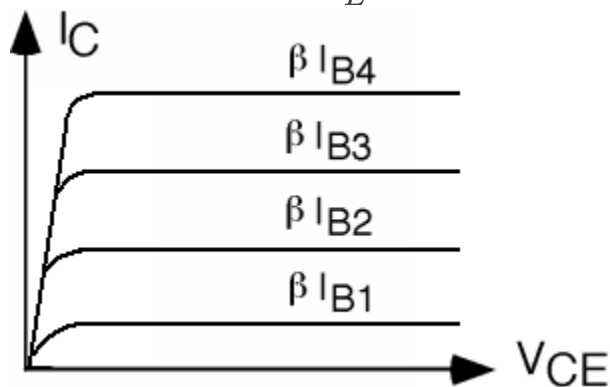
The only drawback with what we have so far is that except in some specialized high-frequency circuits, the bipolar transistor is very rarely used in the common base configuration. Most of the time, you will see it in either [the common emitter configuration](#), or the common collector configuration. The common emitter is probably the way the transistor is most often used.



Configuration for the common

emitter circuit

Note that we have a current source driving the base, and we have applied just one battery all the way from the collector to the emitter. The battery now has to do two things: a) It has to provide reverse bias for the base-collector junction and b) it has to provide forward bias for the base-emitter junction. For this reason, the I_C as a function of V_{CE} curves look a little different now. It is now necessary for V_{CE} to become slightly positive in order to get the transistor into its active mode. The other difference, of course, is that the collector current is now shown as being βI_B the base current instead of αI_E the emitter current.



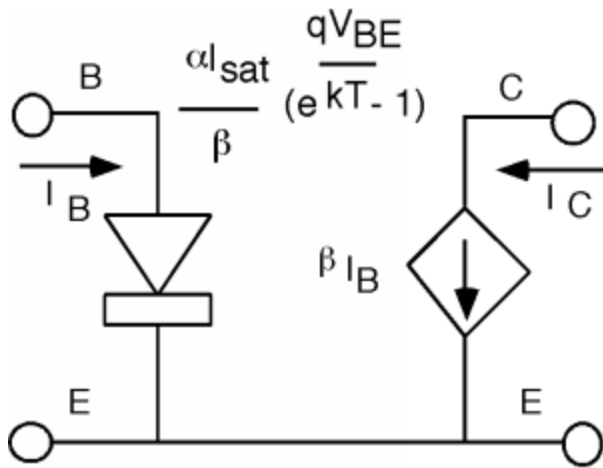
Common emitter characteristic
curves for the transistor

Common Emitter Models

Let's go ahead and draw a model for the transistor in the [common emitter configuration](#). We again have a diode connected between the base and the emitter, and a new current controlled current source between collector and emitter. There is one small caveat which we need to keep in mind however when drawing the common emitter circuit. The diode we see in the base circuit is **not** the same one as we had in the common base model. In the common base model, it was true that

Equation:

$$I_E = I_{\text{sat}} \left(e^{\frac{qV_{BE}}{kT}} - 1 \right)$$



Discrete model for the common emitter configuration

For the base however, only a small fraction of the current that goes through this "diode" actually goes in through the base, the rest is coming in through the collector. Thus we have to make a couple of changes

Equation:

$$\begin{aligned} I_C &= \alpha I_E \\ &= \alpha I_{\text{sat}} \left(e^{\frac{qV_{BE}}{kT}} - 1 \right) \end{aligned}$$

Equation:

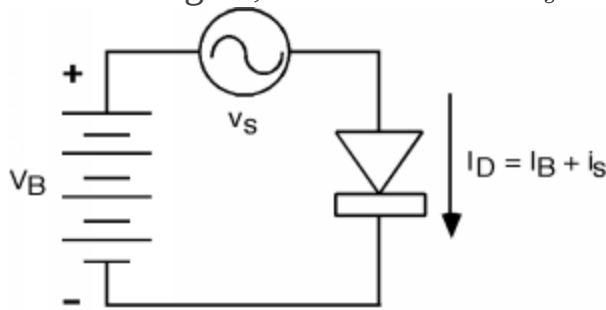
$$\begin{aligned} I_B &= \frac{I_C}{\beta} \\ &= \frac{\alpha I_{\text{sat}}}{\beta} \left(e^{\frac{qV_{\text{BE}}}{kT}} - 1 \right) \end{aligned}$$

So the operational equation for the diode in the base circuit still is the usual exponential function of V_{BE} , except that it now has a saturation current of $\frac{\alpha I_{\text{sat}}}{\beta}$ instead of just I_{sat} .

In principle you could put this model into a circuit, and analyze it to find all of the necessary voltages and currents. However, this would not be very convenient. The base-emitter junction is connected by a diode, which as we know, has a very non-linear I-V relationship. It would be nice if we could come up with a **linear** model which, at least over some limited range of inputs, we could use with confidence.

Small Signal Models

In order to do this we need to introduce the concept of **bias**, and **large signal** and **small signal device** behavior. Consider the following circuit, shown in [\[link\]](#). We are applying the sum of two voltages to the diode, V_B , the **bias voltage** (which is assumed to be a DC voltage) and v_s the **signal voltage** (which is assumed to be AC, or sinusoidal). By definition, we will assume that $|v_s|$ is much less than $|V_B|$. As a result of these voltages, there will be a current I_B flowing through the diode which will consist of two currents, I_B the so-called **bias current**, and i_s , which will be the **signal current**. Again, we assume that i_s is much smaller than I_B .



Putting together a large signal
bias, and a small signal AC
excitation

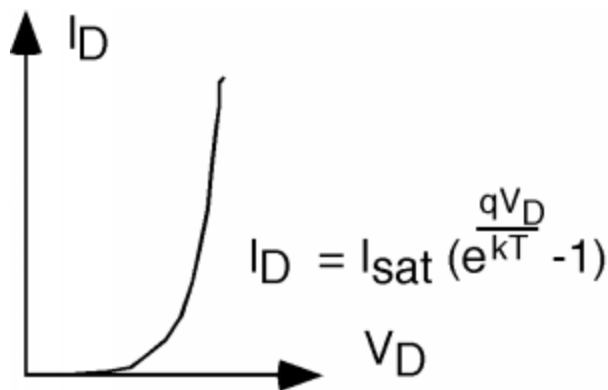
What we would like to do is to see if we can find a linear relationship between v_s and i_s which we could use in our signal analysis. There are two ways we can attack the problem; a graphical approach, and a purely mathematical approach. Let's try the graphical approach first, as it is more intuitive, and then we will confirm what we find out with a mathematical one.

Let's remind ourselves about the I-V characteristics of a diode. In the present situation, V_D is the sum of two voltages, a DC bias voltage V_B and an AC signal, v_s . Let's plot $V_D(t)$ on the V_D axis as shown in [\[link\]](#). How are we going to figure out what the current is? What we need to do is to

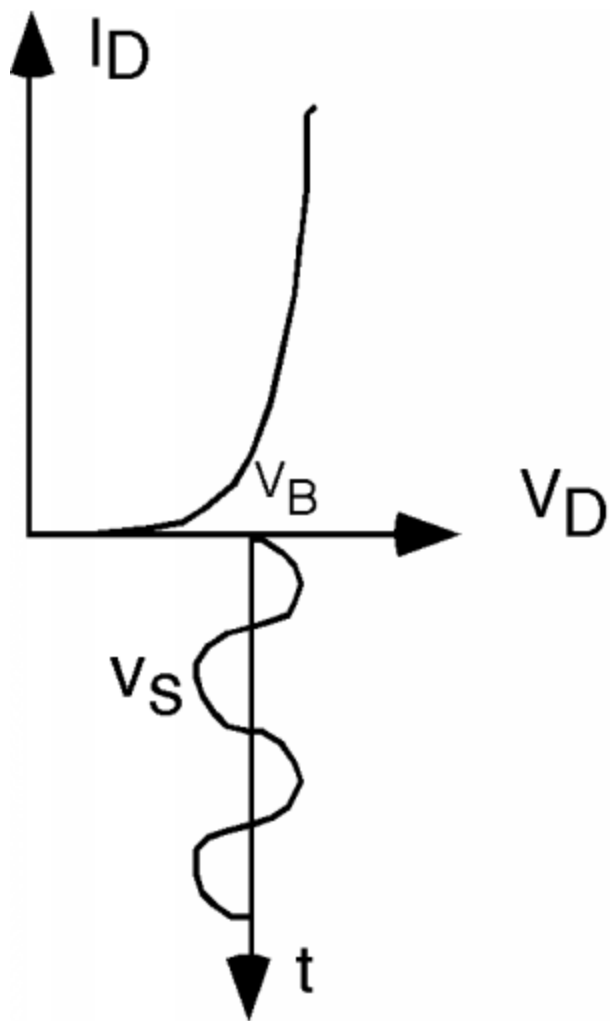
project the voltage up onto the characteristic I-V curve, and then project over to the vertical current axis. We do this in [\[link\]](#). Note that the output current signal is somewhat distorted, which means we do not have linear behavior yet. Let's reduce the amplitude of the signal voltage, as shown in [\[link\]](#). Now we see two things: a) the output is much less distorted, so we must get a more linear behavior, and b) we could get the amplitude of the output signal i_s simply by multiplying the input signal v_s by the slope of the I-V curve at the point where the device is biased. **We have replaced the non-linear I-V curve of the diode by a linear one, which is applicable over the range of the signal voltage.**

Equation:

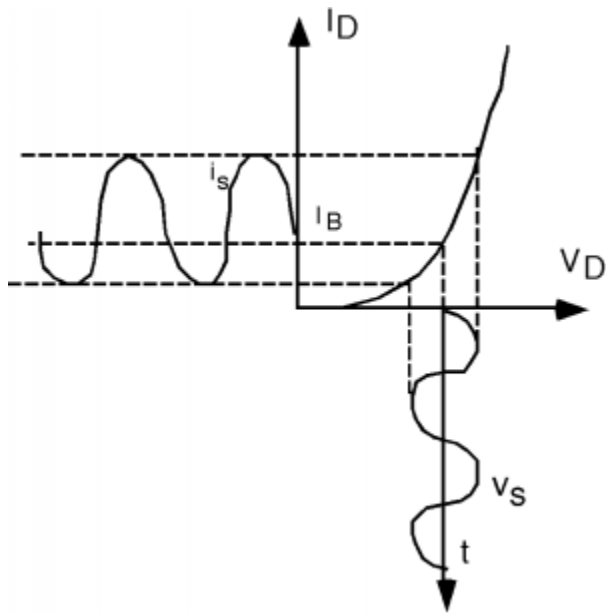
$$i_s = \left. \frac{d}{d V_D} (I_D) \right|_{I_D=I_B}$$



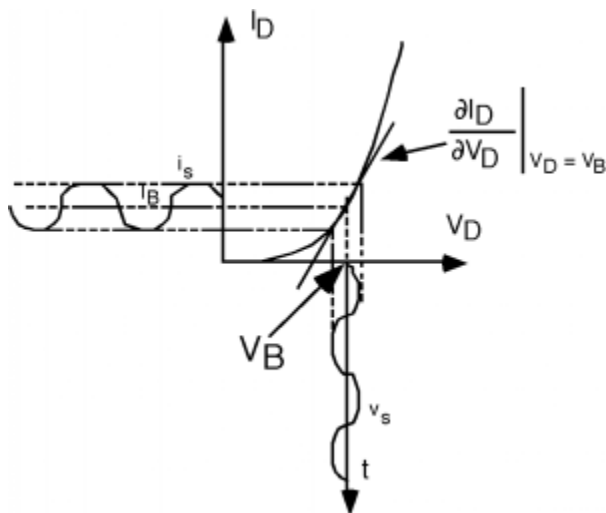
Diode I-V behavior



Bias and signal excitation of a diode I-V curve



Graphically finding the AC response



With a smaller signal, the response is more linear

To get the slope, we need a few simple equations:

Equation:

$$I_D = I_{\text{sat}} \left(e^{\frac{qV_D}{kT}} - 1 \right) \simeq I_{\text{sat}} e^{\frac{qV_D}{kT}}$$

Equation:

$$\frac{d}{d V_D} (I_D) = \frac{q}{kT} I_{\text{sat}} e^{\frac{qV_D}{kT}}$$

When we evaluate the partial derivative at voltage V_D , we note that

Equation:

$$I_{\text{sat}} e^{\frac{qV_D}{kT}} = I_B$$

and hence, the slope of the curve is just $\frac{q}{kT} I_B$ or $40I_B$, since $\frac{q}{kT}$ just has a value of $40V^{-1}$ at room temperatures. Note that current divided by voltage is just conductance, (which is just the inverse of resistance) and so we have found the **small signal linear conductance for the diode**.

As far as the AC signal generator is concerned, we could replace the diode with a resistor whose value is the inverse of the conductance, or $r = \frac{1}{40} I_B$, where I_B is the DC bias current through the diode.

Students are sometimes confused about how we can replace a diode, which only conducts in one direction, with a resistor, which conducts both ways. The answer is to look carefully at [\[link\]](#). As the AC signal voltage rises and falls, the AC output current also increases and decreases in the same manner. Over the limited range of the AC signal parameters, the diode is indeed a linear signal element, not a rectifying one, as it is for large signal applications.

Now let's get the same answer from a purely mathematical approach.

Equation:

$$I_D = I_B + i_s = I_{\text{sat}} \left(e^{\frac{qV_D}{kT}} - 1 \right) \simeq e^{\frac{q(V_B + v_s)}{kT}}$$

In the last expression, we dropped the -1 as it is very small compared to the exponential term and can be neglected.

Now we note that:

Equation:

$$e^{\frac{q(V_B+v_s)}{kT}} = e^{\frac{qV_B}{kT}} e^{\frac{qv_s}{kT}}$$

And, for the second exponential, if qV_B is much less than kT ,

Equation:

$$e^{\frac{qv_s}{kT}} \simeq 1 + \frac{qv_s}{kT} + \dots$$

where we have used the power series expansion for the exponential, but have only taken the first two terms. Thus

Equation:

$$I_B + i_s \simeq I_{\text{sat}} e^{\frac{qV_B}{kT}} \left(1 + \frac{qv_s}{kT} \right)$$

Obviously

Equation:

$$I_B = I_{\text{sat}} e^{\frac{qV_B}{kT}}$$

and

Equation:

$$\begin{aligned} i_s &= I_{\text{sat}} e^{\frac{qV_B}{kT}} \left(\frac{q}{kT} v_s \right) \\ &= \frac{q}{kT} I_B v_s \end{aligned}$$

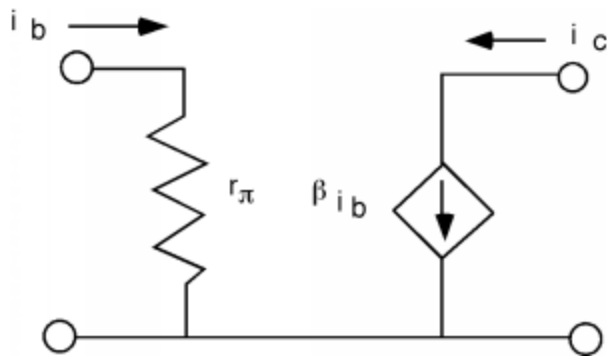
which gives us the same result as before

Equation:

$$\frac{i_s}{v_s} = \frac{q}{kT} I_B$$

Small Signal Model for Bipolar Transistor

Thus if we go back to [the circuit model](#) for the common emitter transistor, and re-draw it as a **small signal model** it would look something like [\[link\]](#). Here we have replaced the diode with a linear element (a resistor, called r_π) and we have changed the notation for the currents from I_B and I_C to i_b and i_c respectively, to remind us that we are now talking about small signal ac quantities, not large signal ones. The bias currents I_B and I_C are still flowing through the device (and we will leave it to ELEC 342 to discuss how these are generated and set up) but they do not appear in the small signal model. This model is only used to figure out how the transistor behaves for the ac signal going through it, not how it responds to large DC values.



Small signal linear model for the common emitter transistor

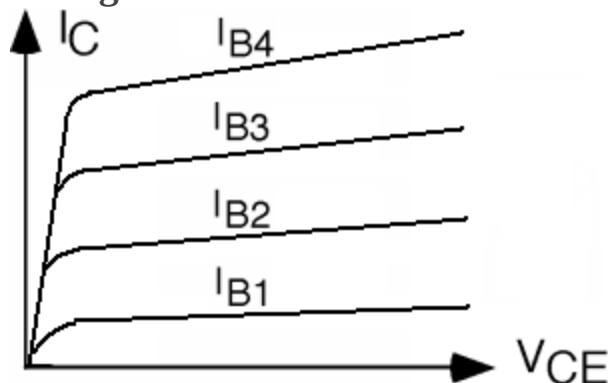
Now r_π the equivalent small signal resistance of the base-emitter diode is given simply by the inverse of the conductance of the equivalent diode. Remember, we found

Equation:

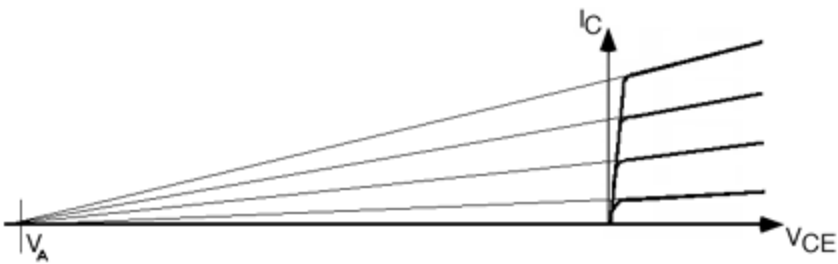
$$\begin{aligned} r_\pi &= \frac{1}{\frac{q}{kT} I_B} \\ &= \frac{1}{\frac{q}{kT} \frac{I_C}{\beta}} \\ &= \frac{\beta}{40 I_C} \end{aligned}$$

where we have used the fact that $I_C = \beta I_B$ and $\frac{q}{kT} = 40V^{-1}$. As we said earlier, typical values for β in a standard bipolar transistor will be around 100. Thus, for a typical collector bias current of $I_C = 1\text{ mA}$, r_π will be about $2.5\text{ k}\Omega$.

There is one more item we should consider in putting together our model for the bipolar transistor. We did not get things completely right when we drew the [common emitter characteristic curves](#) for the transistor. There is a somewhat subtle effect going on when V_{CE} is increased. Remember, we said that the current coming out of the collector is not effected by how big the drop was in the reverse biased base-collector junction. The collector current just depends on how many electrons are injected into the base by the emitter, and how many of them make it across the base to the base-collector junction. As the base-collector reverse bias is increased (by increasing V_{CE} the depletion width of the base-collector junction increases as well. This has the effect of making the base region somewhat shorter. This means that a few more electrons are able to make it across the base region without recombining and as a result α and hence β increase somewhat. This then means that I_C goes up slightly with increasing V_{CE} . The effect is called **base width modulation**. Let us now include that effect in the common emitter characteristic curves. As you can see in [\[link\]](#), there is now a slope to the $I_C(V_{CE})$ curve, with I_C increasing somewhat as V_{CE} increases. The effect has been somewhat exaggerated in [\[link\]](#), and I will now make the slope even bigger so that we may define a new quantity, called the **Early Voltage**.



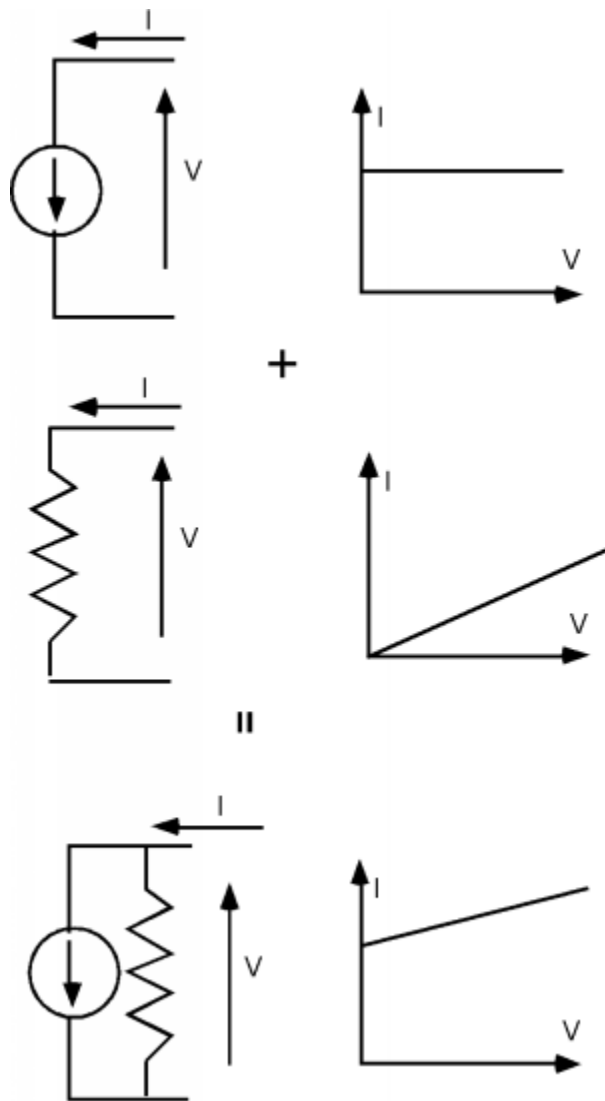
Common emitter response with
base-width modulation effect



Finding the Early Voltage

Back in the very beginning of the transistor era, an engineer at Bell Labs, Jim Early, predicted that there would be a slope to the I_C curves, and that they would all project back to the same intersection point on the horizontal axis. Having made that prediction, Jim went down into the lab, made the measurement, and confirmed his prediction, thus showing that the theory of transistor behavior was being properly understood. The point of intersection of the V_{CE} axis is known as the **Early Voltage**. Since the symbol V_E , for the emitter voltage was already taken, they had to label the Early Voltage V_A instead. (Even though the intersection point is on the negative half of the V_{CE} axis, V_A is universally quoted as a positive number.)

How can we model the sloping I-V curve? We can do almost the same thing as we did with the solar cell. The horizontal part of the curve is still a current source, and the sloped part is simply a resistor in parallel with it. Here is a graphical explanation in [\[link\]](#).



Combining a current source and
a resistor in parallel

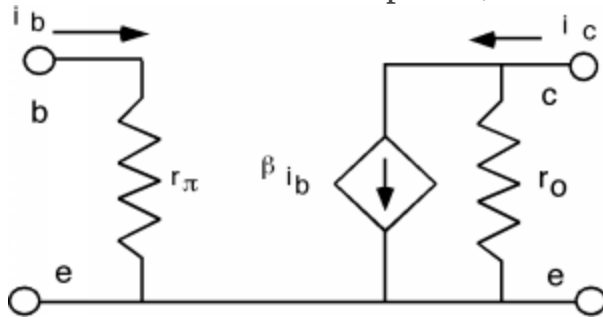
Usually, the slope is much less than we have shown here, and so for any given value of I_C , we can just take the slope of the line as $\frac{I_C}{V_A}$, and hence the resistance, which is usually called r_o is just $\frac{V_A}{I_C}$. Thus, we add r_o to the small signal model for the bipolar transistor. This is shown in [\[link\]](#). In a good quality modern transistor, the Early Voltage, V_A will be on the order of

150-250 Volts. So if we let $V_A = 200$, and we imagine that we have our transistor biased at 1 mA, then

Equation:

$$\begin{aligned} r_o &= \frac{200V}{1 \text{ mA}} \\ &= 200\text{k}\Omega \end{aligned}$$

which is usually much larger than most of the other resistors you will encounter in a typical circuit. In most instances, r_o can be ignored with no problem. If you get into high impedance circuits however, as you might find in a instrumentation amplifier, then v_{be} has to be taken into account.



Including r_o in the small signal
linear model

Sometimes it is advantageous to use a mutual transconductance model instead of a current gain model for the transistor. If we call the input small signal voltage v_{be} , then obviously

Equation:

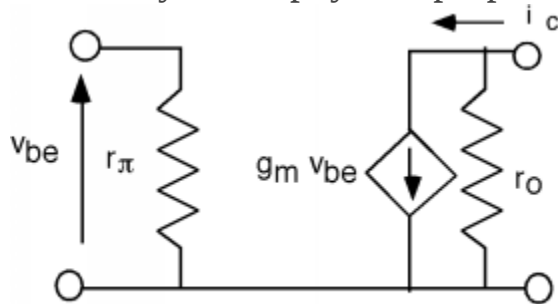
$$\begin{aligned} i_b &= \frac{v_{be}}{r_\pi} \\ &= \frac{v_{be}}{\frac{\beta}{40I_C}} \end{aligned}$$

But

Equation:

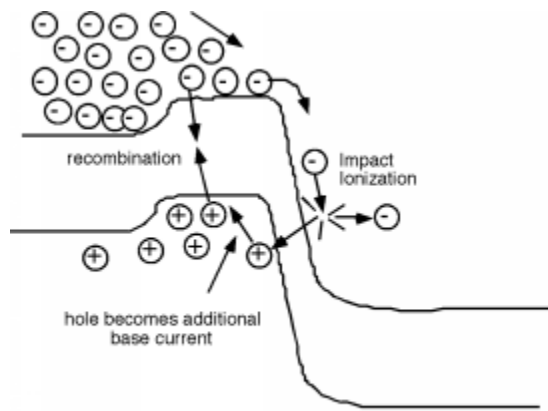
$$i_c = \beta i_b = \frac{\beta v_{be}}{\frac{\beta}{40I_C}} = 40I_C v_{be} \equiv g_m v_{be}$$

Where g_m is called the mutual transconductance of the transistor. Notice that β has completely cancelled out in the expression for g_m and that g_m depends only upon the bias current, I_C , flowing through the collector and not on any of the physical properties of the transistor itself!



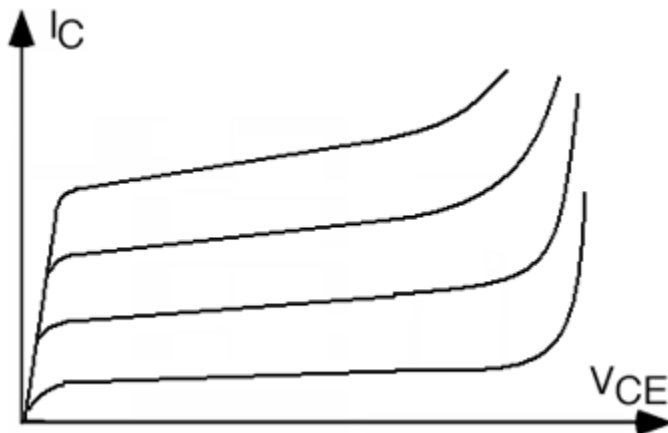
Transconductance small signal
linear model

Finally, there is one last physical consideration we should make concerning the operation of the bipolar transistor. The base-collector junction is reverse biased. We know that if we apply too much reverse bias to a pn junction, it can breakdown through avalanche multiplication. Breakdown in a transistor is somewhat "softer" than for a simple diode, because once a small amount of avalanche multiplication starts, extra holes are generated within the base-collector junction. These holes fall up, into the base, where they act as additional base current, which, in turn, causes I_C to increase. This is shown in [\[link\]](#).



Ionization at the base-collector junction causes additional base current

A set of characteristic curves for a transistor going into breakdown is also shown in [\[link\]](#).



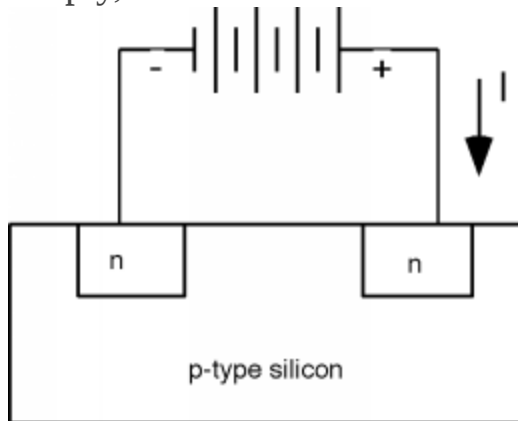
Bipolar Transistor going into breakdown

Well, we have learned quite a bit about bipolar transistors in a very short space. Go back over this chapter and see if you can pick out the two or three

most important ideas of equations which would make up a set of "facts" that you could stick away in you head someplace. Do this so you will always have them to refer to when the subject of bipolars comes up (In say, a job interview or something!).

Introduction to MOSFETs

We now move on to another three terminal device - also called a **transistor**. (In truth this device really has at least four, and probably five, terminals, but we will leave the subtle details for a later time.) This transistor, however, works on much different principles than does the bipolar junction transistor of the last chapter. We will now focus on a device called the **Field Effect Transistor**, or **Metal-Oxide-Semiconductor Field Effect Transistor** or simply, the **MOSFET**. Consider the following:



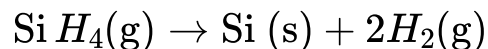
The start of a field effect transistor

Here we have a block of silicon, doped p-type. Into it we have made two regions which are doped n-type. To each of those n-type regions we attach a wire, and connect a battery between them. If we try to get some current, I , to flow through this structure, nothing will happen, because the n-p junction on the RHS is reverse biased (We have the positive lead from the battery going to the n-side of the p-n junction). If we attempt to remedy this by turning the battery around, we will now have the LHS junction reverse biased, and again, no current will flow. If, for whatever reason, we want current to flow, we will need to come up with some way of forming a layer of n-type material between one n-region and the other. This will then

connect them together, and we can run current in one terminal and out the other.

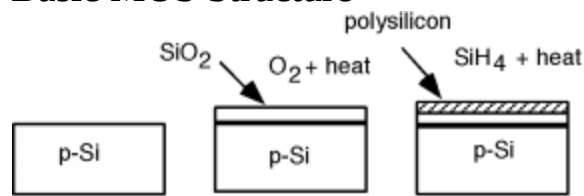
To see how we will do this, let's do two things. First we will grow a layer of SiO_2 (silicon dioxide, or just plain "oxide") on top of the silicon. (This turns out to be relatively easy, we just stick the wafer in an oven with some oxygen flowing through it, and heat everything up to about 1100°C for an hour or so, and we end up with a nice, high-quality insulating SiO_2 layer on top of the silicon). On top of the oxide layer we then deposit a conductor, which we call the gate. In the "old days" the gate would have been a layer of aluminum (Hence the "metal-oxide-silicon" or MOS name). Today, it is much more likely that a heavily doped layer of polycrystalline silicon (polysilicon, or more often just "poly") would be deposited to form the gate structure. (I guess "POS" sounded funny to people in the field, because it never caught on as a name for these devices). Polysilicon is made from the reduction of a gas, such as silane (SiH_4) through the reaction

Equation:



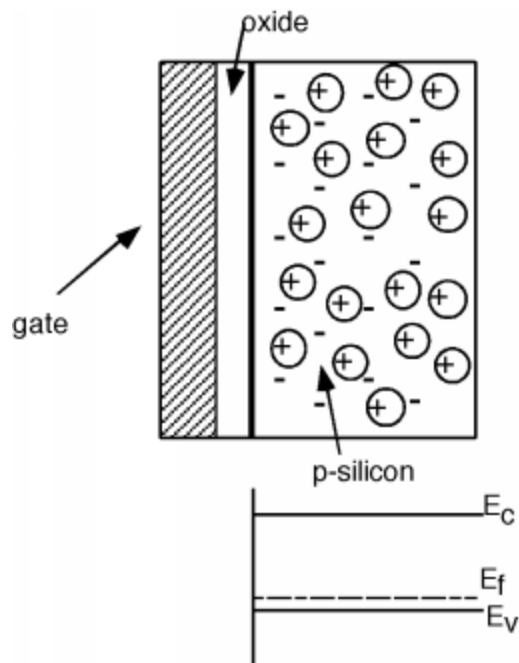
The silicon is polycrystalline (composed of lots of small silicon crystallites) because it is deposited on top of the oxide, which is amorphous, and so it does not provide a single crystal "matrix" which would allow the silicon to organize itself into one single crystal. If we had deposited the silicon on top of a single crystal silicon wafer, we would have formed a single crystal layer of silicon called an **epitaxial layer**. (**Epitaxy** comes from the Greek, and it just means "ordered upon". Thus an epitaxial layer is one which follows the order of the substrate on which it is grown). This is sometimes done to make structures for particular applications. For instance, growing a n-type epitaxial layer on top of a p-type substrate permits the fabrication of a very abrupt p-n junction.

Basic MOS Structure



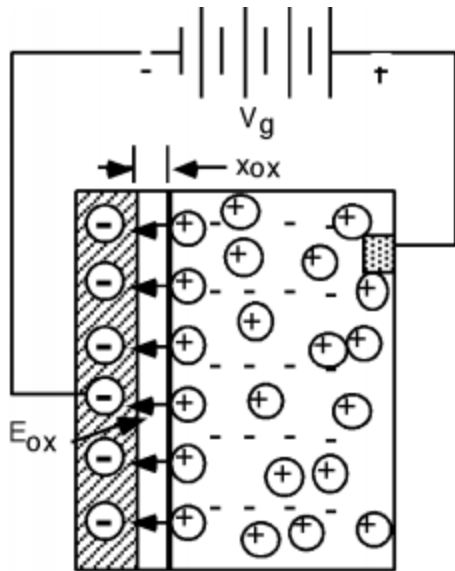
Formation of the MOS structure

[\[link\]](#) shows the steps necessary to make the MOS structure. It will help us in our understanding if we now rotate our picture so that it is pointing sideways in our next few drawings. (Also, we will forget about the two n-regions for awhile, and pick them back up later when we rotate the structure right side up again.) [\[link\]](#) shows the rotated structure. Note that in the p-silicon we have positively charged mobile holes, and negatively charged, fixed acceptors. Because we will need it later, we have also shown the band diagram for the semiconductor below the sketch of the device. Note that since the substrate is p-type, the Fermi level is located down close to the valance band.



Basic MOS structure

Let us now place a potential between the gate and the silicon substrate. Suppose we make the gate negative with respect to the substrate. Since the substrate is p-type, it has a lot of mobile, positively charged holes in it. Some of them will be attracted to the negative charge on the gate, and move over to the surface of the substrate. This is also reflected in the band diagram below the [sketch of the structure](#). Remember that the density of holes is exponentially proportional to how close the Fermi level is to the valence band edge. We see that the band diagram has been bent up slightly near the surface to reflect the extra holes which have accumulated there.



Applying a negative gate
voltage

An electric field will develop between the positive holes and the negative gate charge. Note that the gate and the substrate form a kind of parallel plate capacitor, with the oxide acting as the insulating layer in-between them. The oxide is quite thin compared to the area of the device, and so it is quite appropriate to assume that the electric field inside the oxide is a uniform one. (We will ignore fringing at the edges.) The integral of the electric field is just the applied gate voltage V_g . If the oxide has a thickness x_{ox} then since E_{ox} is uniform, it is given by

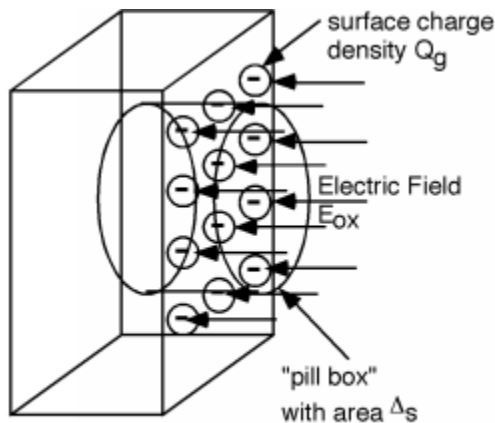
Equation:

$$E_{\text{ox}} = \frac{V_g}{x_{\text{ox}}}$$

If we focus in on a small part of the gate, we can make a little "pill" box which extends from somewhere in the oxide, across the oxide/gate interface and ends up inside the gate material someplace. The pill-box will have an area $\Delta(s)$. Now we will invoke Gauss' law which we reviewed earlier. Gauss' law simply says that the surface integral over a closed surface of the displacement vector D (which is, of course, just ϵ times E) is equal to the total charge enclosed by that surface. We will assume that there is a surface charge density $-Q_g$ ($\frac{\text{Coulombs}}{\text{cm}^2}$) on the [surface of the gate electrode](#). The integral form of Gauss' Law is just:

Equation:

$$\oint \epsilon_{\text{ox}} \mathbf{E} \, d\mathbf{S} = Q_{\text{encl}}$$



Finding the surface charge density

Note that we have used $\epsilon_{\text{ox}}E$ in place of D . In this particular set-up the integral is easy to perform, since the electric field is uniform, and only pointing in through one surface - it terminates on the negative surface charge inside the pill-box. The charge enclosed in the pill box is just $-(Q_g\Delta(s))$, and so we have (keeping in mind that the surface integral of a vector pointing into the surface is negative)

Equation:

$$\begin{aligned}\oint \epsilon_{\text{ox}} \mathbf{E} \, d\mathbf{S} &= -(\epsilon_{\text{ox}} E_{\text{ox}} \Delta(s)) \\ &= -(Q_g \Delta(s))\end{aligned}$$

or

Equation:

$$\epsilon_{\text{ox}} E_{\text{ox}} = Q_g$$

Now, we can use [\[link\]](#) to get

Equation:

$$\frac{\epsilon_{\text{ox}} V_g}{x_{\text{ox}}} = Q_g$$

or

Equation:

$$\frac{Q_g}{V_g} = \frac{\epsilon_{\text{ox}}}{x_{\text{ox}}} \equiv c_{\text{ox}}$$

The quantity c_{ox} is called the **oxide capacitance**. It has units of $\frac{\text{Farads}}{\text{cm}^2}$, so it is really a capacitance **per unit area** of the oxide. The dielectric constant of silicon dioxide, ϵ_{ox} , is about $3.3 \times 10^{-13} \text{F/cm}$. A typical oxide thickness might be 250 \AA (or $2.5 \times 10^{-6} \text{cm}$). In this case, c_{ox} would be about $1.30 \times 10^{-7} \frac{\text{F}}{\text{cm}^2}$. (The units we are using here, while they might seem a little arbitrary and confusing, are the ones most commonly used in the semiconductor business. You will get used to them in a short while.)

The most useful form of [\[link\]](#) is when it is turned around:

Equation:

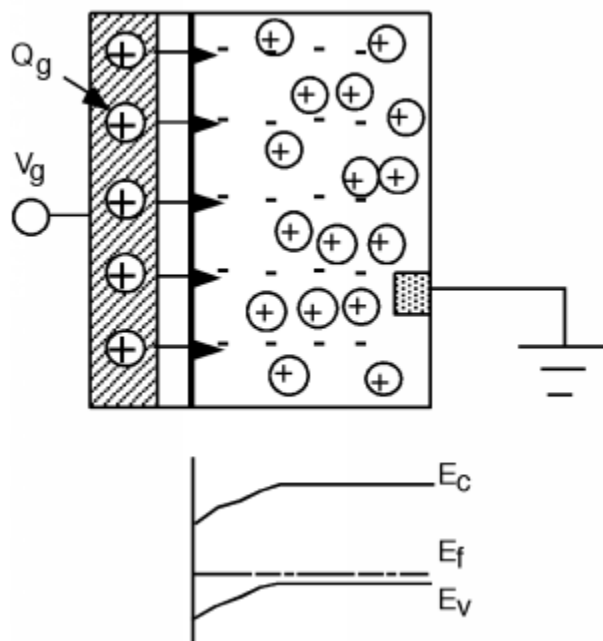
$$Q_g = c_{\text{ox}} V_g$$

as it gives us a way to find the charge on the gate in terms of the gate potential. We will use this equation later in our development of how the MOS transistor really works.

It turns out we have not done anything very useful by apply a negative voltage to the gate. We have drawn more holes there in what is called an **accumulation layer**, but that is not helping us in our effort to create a layer of electrons in the MOSFET which could electrically connect the two n-regions together.

Let's turn the battery around and apply a **positive** voltage to the gate. (Actually, let's take the battery out of the [sketch](#) for now, and just let V_g be a

positive value, relative to the substrate which will tie to ground.) Making V_g positive puts positive Q_g on the gate. The positive charge pushes the holes away from the region under the gate and uncovers some of the negatively-charged fixed acceptors. Now the electric field points the other way, and goes from the positive gate charge, terminating on the negative acceptor charge within the silicon.



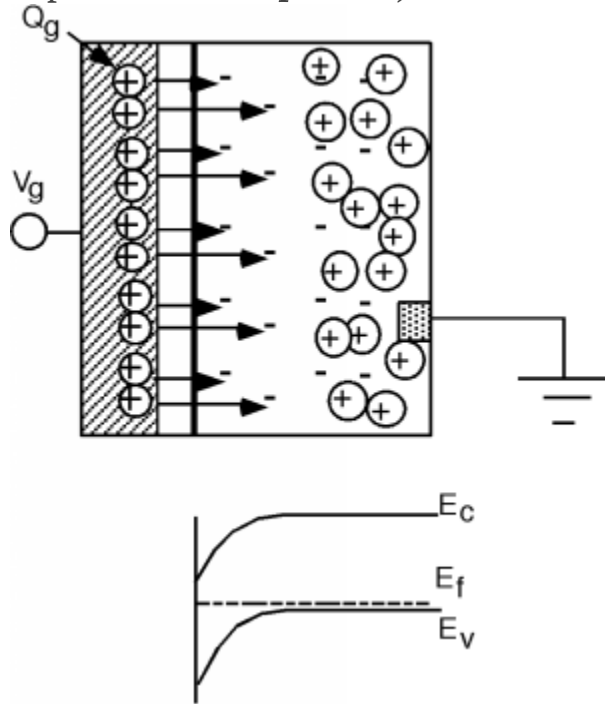
Increasing the voltage extends the depletion region further into the device

The electric field now extends **into** the semiconductor. We know from our experience with the p-n junction that when there is an electric field, there is a shift in potential, which is represented in the band diagram by bending the bands. Bending the bands down (as we should moving towards positive charge) causes the valence band to pull away from the Fermi level near the surface of the semiconductor. If you remember the expression we had for the density of holes in terms of E_v and E_f ([electron and hole density equations](#)) it is easy to see that indeed

Equation:

$$p = N_v e^{-\frac{E_f - E_v}{kT}}$$

there is a depletion region (region with almost no holes) near the region under the gate. (Once $E_f - E_v$ gets large with respect to kT , the negative exponent causes $p \rightarrow 0$.)



Threshold, E_f is getting close to E_c

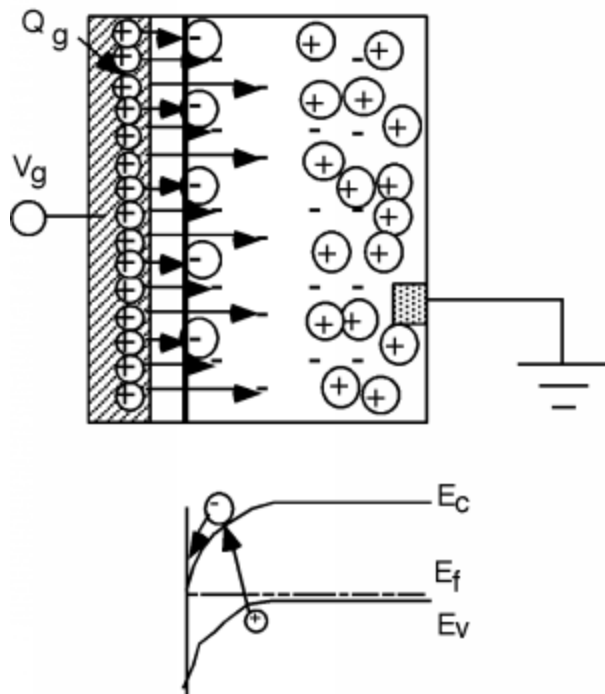
The electric field extends further into the semiconductor, as more negative charge is uncovered and the bands bend further down. But now we have to recall the [electron density equation](#), which tells us how many electrons we have

Equation:

$$n = N_c e^{-\frac{E_c - E_f}{kT}}$$

A glance at [\[link\]](#) above reveals that with this much band bending, E_c the conduction band edge, and E_f the Fermi level are starting to get close to one another (at least compared to kT), which means that n , the electron concentration, should soon start to become significant. In the situation represented by [\[link\]](#), we say we are at **threshold**, and the gate voltage at this point is called the **threshold voltage**, V_T .

Now, let's increase V_g above V_T . Here's the sketch in [\[link\]](#).



Inversion - Electrons form an inversion layer under the gate

Even though we have increased V_g beyond the threshold voltage, V_T , and more positive charge appears on the gate, the depletion region no longer moves back into the substrate. Instead electrons start to appear under the gate region, and the additional electric field lines terminate on these new electrons, instead of on additional acceptors. We have created an **inversion layer** of electrons under the gate, and it is this layer of electrons which we can use to connect the two n-type regions in our initial device.

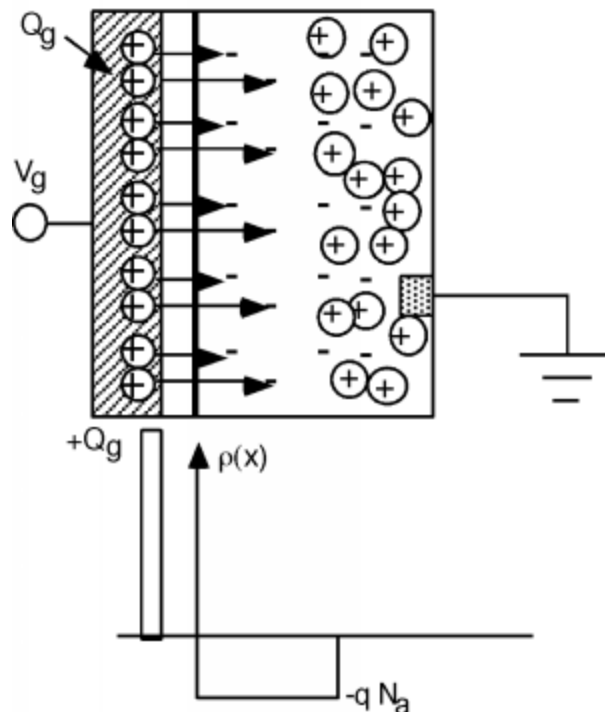
Where did these electrons come from? We do not have any donors in this material, so they can not come from there. The only place from which electrons could be found would be through thermal generation. Remember, in a semiconductor, there are always a few electron hole pairs being generated by thermal excitation at any given time. Electrons that get created in the depletion region are caught by the electric field and are swept over to the edge by the gate. I have tried to suggest this with the electron generation event shown in the band diagram in the figure. In a **real** MOS device, we have the two n-regions, and it is easy for electrons from one or both to "fall" into the potential well under the gate, and create the inversion layer of electrons.

Threshold Voltage

Our task now is to figure out how much voltage we need to get V_g up to V_T and then to figure out how much negative charge there is under the gate, once V_T has been exceeded. The first part is actually pretty easy. It is a lot like the problem we looked at, with the one-sided diode, but with just a little added complication. To start out, let's make a sketch of the charge density distribution under the conditions of [this image](#), just when we get to threshold. We'll include the sketch of the structure too, so it will be clear what charge we are talking about. This is shown in [\[link\]](#). Now, we just use the equation we developed before for the [electric field](#), which came from integrating the differential form of Gauss' Law.

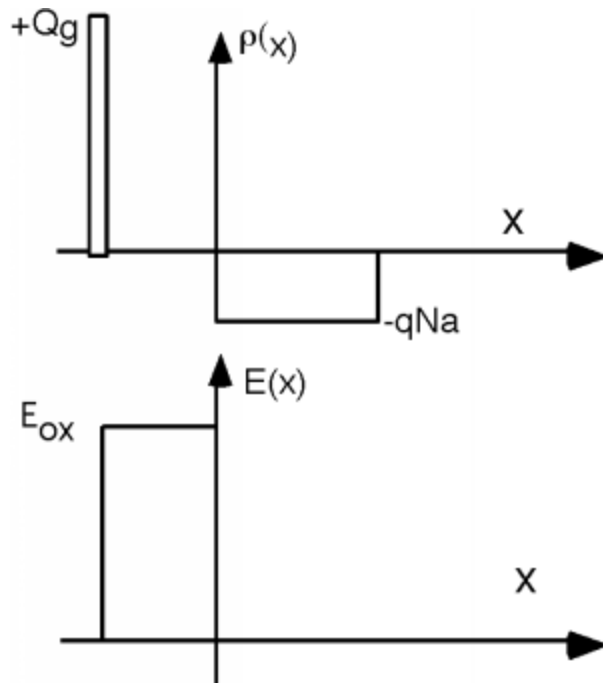
Equation:

$$E(x) = \int \frac{\rho(x)}{\epsilon} dx$$



Charge distribution at threshold

As before, we will do the integral graphically, starting at the LHS of the picture. The field outside the structure must be zero, so we have no electric field until we get to the delta function of charge on the gate, at which time it jumps up to some value we will call E_{ox} . There is no charge inside the oxide, so $\frac{d}{dx}(E)$ is zero, and thus $E(x)$ must remain constant at E_{ox} until we reach the oxide/silicon interface.



Electric field in the oxide

If we were to put our little "pill box" on the oxide-silicon interface, the integral of D over the face in the silicon would be $\epsilon_{Si}E_{Si}\Delta(S)$, where E_{Si} is the strength of the electric field inside the silicon. On the face inside the oxide it would be $-(\epsilon_{ox}E_{ox}\Delta(S))$, where E_{ox} is the strength of the electric field in the oxide. The minus sign comes from the fact that the field on the oxide side is going into the pill box instead of out of it. There is no net charge contained within the pill box, so the sum of these two integrals must be zero. (The integral over the **entire** surface equals the enclosed charge, which is zero.)

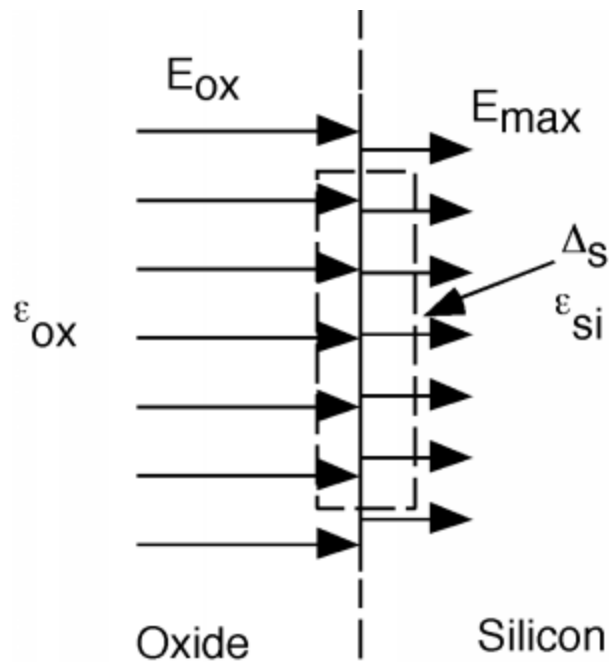
Equation:

$$\epsilon_{\text{Si}} E_{\text{max}} \Delta(S) - \epsilon_{\text{ox}} E_{\text{ox}} \Delta(S) = 0$$

or

Equation:

$$\epsilon_{\text{Si}} E_{\text{max}} = \epsilon_{\text{ox}} E_{\text{ox}}$$



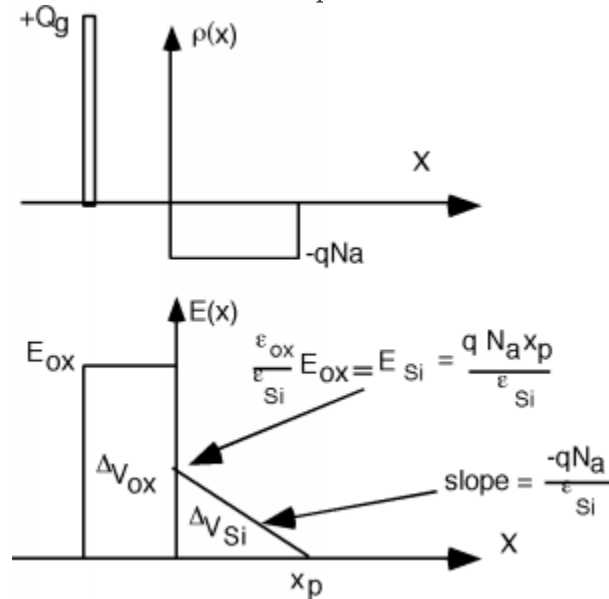
Using Gauss' Law at the
silicon/oxide interface

This is just a statement that it is the normal component of displacement vector, D , which must be continuous across a dielectric interface, not the electric field, E . Solving [\[link\]](#) for the electric field in the silicon:

Equation:

$$E_{\text{Si}} = \frac{\epsilon_{\text{ox}}}{\epsilon_{\text{Si}}} E_{\text{ox}}$$

The dielectric constant of oxides about one third that of the dielectric constant of silicon dioxide, so we see a "jump" down in the magnitude of the electric field as we go from oxide to silicon. The charge density in the depletion region of the silicon is just $-(qN_a)$ and so the electric field now starts decreasing at a rate $\frac{-(qN_a)}{\epsilon_{Si}}$ and reaches zero at the end of the depletion region, x_p .



electric field and voltage drops
across the entire structure

Clearly, we have two different regions, each with their own voltage drop. (Remember the integral of electric field is voltage, so the area under each region of $E(x)$ represents a voltage drop.) The drop in the little triangular region we will call $\Delta(V_{Si})$ and it represents the potential drop in going from the bulk, down to the bottom of the drooping conduction band at the silicon-oxide interface. Looking back at [the earlier figure](#) on threshold, you should be able to see that this is nearly one whole band-gaps worth of potential, and so we can safely say that $(\Delta(V_{Si}) \simeq 0.8) \rightarrow 1.0V$.

Just as with the single-sided diode, the width of the depletion region x_p , is (which we saw in [a previous equation](#)):

Equation:

$$x_p = \sqrt{\frac{2\varepsilon_{\text{Si}}\Delta(V_{\text{Si}})}{qN_a}}$$

from which we can get an expression for E_{Si}

Equation:

$$\begin{aligned} E_{\text{Si}} &= \frac{qN_a}{\varepsilon_{\text{Si}}} x_p \\ &= \sqrt{\frac{2qN_a\Delta(V_{\text{Si}})}{\varepsilon_{\text{Si}}}} \end{aligned}$$

by multiplying the slope of the $E(x)$ line by the width of the depletion region, x_p .

We can now use [\[link\]](#) to find the electric field in the oxide

Equation:

$$\begin{aligned} E_{\text{ox}} &= \frac{\varepsilon_{\text{Si}}}{\varepsilon_{\text{ox}}} E_{\text{Si}} \\ &= \frac{1}{\varepsilon_{\text{ox}}} \sqrt{2q\varepsilon_{\text{Si}}N_a\Delta(V_{\text{Si}})} \end{aligned}$$

Finally, $\Delta(V_{\text{ox}})$ is simply the product of E_{ox} and the oxide thickness, x_{ox}

Equation:

$$\begin{aligned} \Delta(V_{\text{ox}}) &= x_{\text{ox}} E_{\text{ox}} \\ &= \frac{x_{\text{ox}}}{\varepsilon_{\text{ox}}} \sqrt{2q\varepsilon_{\text{Si}}N_a\Delta(V_{\text{Si}})} \end{aligned}$$

Note that x_{ox} divided by ε_{ox} is simply one over c_{ox} , the oxide capacitance, which we described earlier. Thus

Equation:

$$\Delta(V_{\text{ox}}) = \frac{1}{c_{\text{ox}}} \sqrt{2q\epsilon_{\text{Si}}N_a\Delta(V_{\text{Si}})}$$

And the threshold voltage V_T is then given as

Equation:

$$\begin{aligned} V_T &= \Delta(V_{\text{Si}}) + \Delta(V_{\text{ox}}) \\ &= \Delta(V_{\text{Si}}) + \frac{1}{c_{\text{ox}}} \sqrt{2q\epsilon_{\text{Si}}N_a\Delta(V_{\text{Si}})} \end{aligned}$$

which is not that hard to calculate! [\[link\]](#) is one of the most important equations in this discussion of field effect transistors, as it tells us when the MOS device is turned on.

[\[link\]](#) has several "handles" available to the device engineer to build a device with a given threshold voltage. We know that as we increase N_a , the acceptor density, that the Fermi level gets closer to the valance band, and hence $\Delta(V_{\text{Si}})$ will change some. But as we said, it will always be around 0.8 to 1 Volt, so it will not be the driving term which dominates V_T . Let's see what we get with an acceptor concentration of 10^{17} . Just for completeness, let's calculate $E_f - E_v$.

Equation:

$$\begin{aligned} p &= N_a \\ &= N_v e^{\frac{E_f - E_v}{kT}} \end{aligned}$$

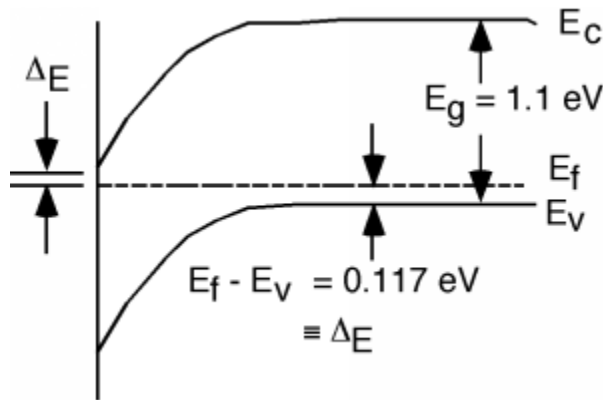
$$\text{thus } E_f - E_v = kT \ln\left(\frac{N_v}{N_a}\right)$$

In silicon, N_v is 1.08×10^{19} and this makes $E_f - E_v = 0.117$ eV which we will call $\Delta(E)$. It is conventional to say that a surface is inverted if, at the silicon surface, $E_c - E_f$, the distance between the conduction band and the Fermi level is the same as the distance between the Fermi level and the valance band in the bulk. With a little time spent looking at [\[link\]](#), you

should be able to convince yourself that the total energy change in going from the bulk to the surface in this case would be

Equation:

$$\begin{aligned}
 q\Delta(V_{\text{Si}}) &= E_g - 2\Delta(E) \\
 &= 1.1 \text{ eV} - 2 \times (0.117 \text{ eV}) \\
 &= 0.866 \text{ eV}
 \end{aligned}$$



Example of finding $\Delta(V_{\text{Si}})$

Using $N_A = 10^{17}$, $\epsilon_{\text{Si}} = 1.1 \times 10^{-12} \frac{\text{F}}{\text{cm}}$ and $q = 1.6 \times 10^{-19} \text{C}$, we find that

Equation:

$$\sqrt{2q\epsilon_{\text{Si}}N_A\Delta(V_{\text{Si}})} = 1.74 \times 10^{-7}$$

We saw earlier that if we have an oxide thickness of 250\AA , we get a value for c_{ox} of $1.3 \times 10^{-7} \frac{\text{F}}{\text{cm}^2}$ ($\frac{\text{Coulombs}}{\text{Vcm}^2}$), and so

Equation:

$$\begin{aligned}
\Delta(V_{\text{ox}}) &= \frac{1}{c_{\text{ox}}} \sqrt{2q\varepsilon_{\text{Si}}N_a\Delta(V_{\text{Si}})} \\
&= \frac{1}{1.3 \times 10^{-7}} 1.74 \times 10^{-7} \\
&= 1.32V
\end{aligned}$$

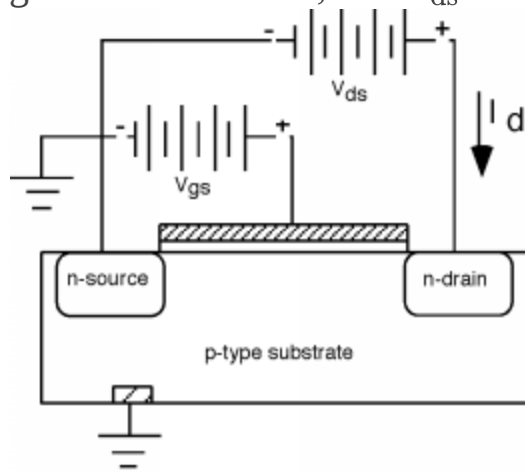
and

Equation:

$$\begin{aligned}
V_T &= \Delta(V_{\text{Si}}) + \Delta(V_{\text{ox}}) \\
&= 0.866 + 1.32 \\
&= 2.18V
\end{aligned}$$

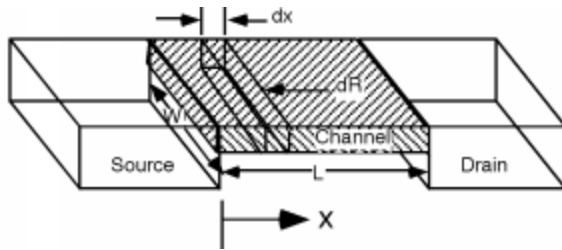
MOS Transistor

Now we can go back now to our initial structure, shown in the [introduction to MOSFETs](#), only this time we will add an oxide layer, a gate structure, and another battery so that we can invert the region under the gate and connect the two n-regions together. We'll also identify some names for parts of the structure, so we will know what we are talking about. For reasons which will be clear later, we call the n-region connected to the negative side of the battery the **source**, and the other one the **drain**. We will ground the source, and also the p-type substrate. We add two batteries, V_{gs} between the gate and the source, and V_{ds} between the drain and the source.



Biassing a MOSFET transistor

It will be helpful if we also make another sketch, which gives us a perspective view of the device. For this we strip off the gate and oxide, but we will imagine that we have applied a voltage greater than V_T to the gate, so there is a n-type region, called the **channel** which connects the two. We will assume that the channel region is L long and W wide, as shown in [\[link\]](#).



The inversion channel and its resistance

Next we want to take a look at a little section of channel, and find its resistance $d(R)$, when the little section is $d(x)$ long.

Equation:

$$d(R) = \frac{dx}{\sigma_s W}$$

We have introduced a **slightly** different form for our resistance formula here. Normally, we would have a simple σ in the denominator, and an area A , for the cross-sectional area of the channel. It turns out to be very hard to figure out what that cross sectional area of the channel is however. The electrons which form the inversion layer crowd into a very thin sheet of **surface charge** which really has little or no thickness, or penetration into the substrate.

If, on the other hand we consider a surface conductivity (units: simply mhos), σ_s , where

Equation:

$$\sigma_s = \mu_s Q_{\text{chan}}$$

then we will have an expression which we can evaluate. Here, μ_s is a surface mobility, with units of $\frac{\text{cm}^2}{\text{V sec}}$. We ran into μ in [earlier chapters](#), when we were building our simple conduction model. It was the quantity which

represented the proportionality between the average carrier velocity and the electric field.

Equation:

$$\bar{v} = \mu E$$

Equation:

$$\mu = \frac{q\tau}{m}$$

The surface mobility is a quantity which has to be measured for a given system, and is usually just a number which is given to you. Something around $300 \frac{\text{cm}^2}{\text{V sec}}$ is about right for silicon. Q_{chan} is called the surface charge density or channel charge density and it has units of $\frac{\text{Coulombs}}{\text{cm}^2}$. This is like a sheet of charge, which is different from the bulk charge density, which has units of $\frac{\text{Coulombs}}{\text{cm}^2}$. Note that:

Equation:

$$\begin{aligned} \frac{\text{cm}^2}{\text{Volt sec}} \frac{\text{Coulombs}}{\text{cm}^2} &= \frac{\frac{\text{Coul}}{\text{sec}}}{\text{Volt}} \\ &= \frac{I}{V} \\ &= \text{mhos} \end{aligned}$$

It turns out that it is pretty simple to get an expression for Q_{chan} , the surface charge density in the channel. For any given gate voltage V_{gs} , [we know](#) that the charge density on the gate is given simply as:

Equation:

$$Q_g = c_{\text{ox}} V_{\text{gs}}$$

However, until the gate voltage V_{gs} gets larger than V_T we are not creating any mobile electrons under the gate, we are just building up a depletion region. We'll define Q_T as the charge on the gate necessary to get to

threshold. $Q_T = c_{\text{ox}} V_T$. Any charge added to the gate above Q_T is matched by charge Q_{chan} in the channel. Thus, it is easy to say:

Equation:

$$Q_{\text{channel}} = Q_g - Q_T$$

or

Equation:

$$Q_{\text{chan}} = c_{\text{ox}} (V_g - V_T)$$

Thus, putting [\[link\]](#) and [\[link\]](#) into [\[link\]](#), we get:

Equation:

$$\mathcal{d}(R) = \frac{\mathcal{d}(x)}{\mu_s c_{\text{ox}} (V_{\text{gs}} - V_T) W}$$

If you look back at [\[link\]](#), you will see that we have defined a current I_d flowing into the drain. That current flows through the channel, and hence through our little incremental resistance $\mathcal{d}(R)$, creating a voltage drop $\mathcal{d}(V_c)$ across it, where V_c is the channel voltage.

Equation:

$$\begin{aligned} \mathcal{d}(V_c(x)) &= I_d \mathcal{d}(R) \\ &= \frac{I_d \mathcal{d}(x)}{\mu_s c_{\text{ox}} (V_{\text{gs}} - V_T) W} \end{aligned}$$

Let's move the denominator to the left, and integrate. We want to do our integral completely along the channel. The voltage on the channel $V_c(x)$ goes from 0 on the left to V_{ds} on the right. At the same time, x is going from 0 to L . Thus our limits of integration will be 0 and V_{ds} for the voltage integral $\mathcal{d}(V_c(x))$ and from 0 to L for the x integral $\mathcal{d}(x)$.

Equation:

$$\int_0^{V_{\text{ds}}} \mu_s c_{\text{ox}} (V_{\text{gs}} - V_T) W \, dV_c = \int_0^L I_d \, dx$$

Both integrals are pretty trivial. Let's swap the equation order, since we usually want I_d as a function of applied voltages.

Equation:

$$I_d L = \mu_s c_{\text{ox}} W (V_{\text{gs}} - V_T) V_{\text{ds}}$$

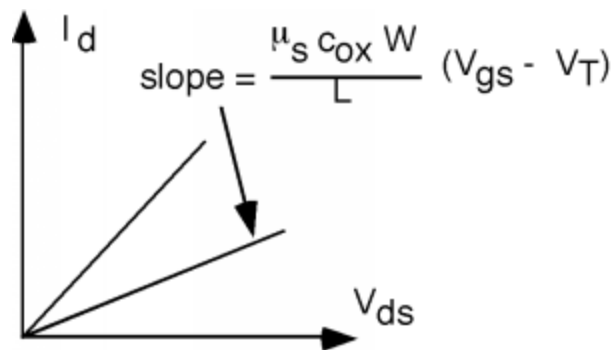
We now simply divide both sides by L , and we end up with an expression for the drain current I_d , in terms of the drain-source voltage, V_{ds} , the gate voltage V_{gs} and some physical attributes of the MOS transistor.

Equation:

$$I_d = \frac{\mu_s c_{\text{ox}} W}{L} (V_{\text{gs}} - V_T) V_{\text{ds}}$$

MOS Regimes

This equation looks a lot like the I-V characteristics of a resistor! I_d is simply proportional to the drain voltage V_{ds} . The proportionality constant depends on the dimensions of the device, W and L as they intuitively should. The current increases as the transistor gets wider, it decreases as it gets longer. It also depends on c_{ox} and μ_s , and on the difference between the gate voltage and the threshold voltage V_T . Note that if we adjust V_{gs} we can change the slope of the I-V curve. We have made a voltage-controlled resistor!



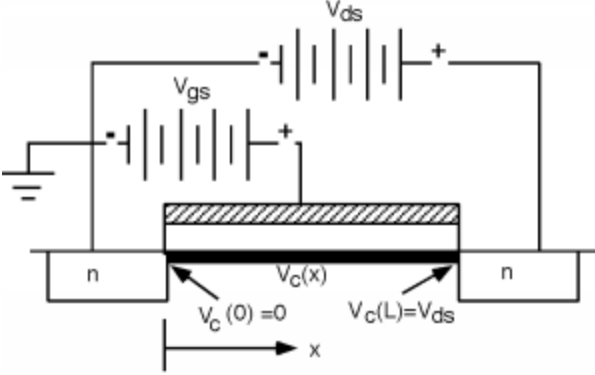
The MOSFET I-V in the linear regime

Caution is advised with this result however, because we have overlooked something quite important. Let's go back to our picture of the gate and the batteries involved in the operation of the MOS transistor. Here we have explicitly shown the channel as a black band and we have introduced a new quantity, $V_c(x)$, the voltage along the channel, and a coordinate x , which tells us where we are on the channel relative to the source and drain. Note that once we apply a drain source potential, V_{ds} , the potential in the channel $V_c(x)$ changes with distance along the channel. At the source end, $V_c(0) = 0$, as the source is grounded. At the drain end, $V_c(L) = V_{ds}$. We will define a voltage V_{gc} which is the potential difference between the gate voltage and the voltage in the channel.

Equation:

$$V_{gc}(x) \equiv V_{gs} - V_c(x)$$

Thus, V_{gc} goes from V_{gs} at the source end to $V_{gs} - V_{ds}$ at the drain end.



Effect of V_{ds} on channel potential

The net charge density in the channel depends upon the potential difference between the **gate and the channel at each point along the channel**, not just $V_{gs} - V_T$. Thus we have to modify [the equation of another module](#) to take this into account

Equation:

$$\begin{aligned} Q_{\text{chan}} &= c_{\text{ox}} (V_{gc}(x) - V_T) \\ &= c_{\text{ox}} (V_{gs} - V_c(x) - V_T) \end{aligned}$$

This, in turn, modifies the [integral relation](#) between I_d and V_{gs} .

Equation:

$$\int_0^{V_{ds}} \mu_s c_{\text{ox}} (V_{gs} - V_T - V_c(x)) W d V_c(x) = \int_0^L I_d d x$$

[\[link\]](#) is only slightly harder to integrate than the one before (Now what is the integral of $x dx$), and so we get for the drain current

Equation:

$$I_d = \frac{\mu_s c_{ox} W}{L} \left((V_{gs} - V_T) V_{ds} - \frac{V_{ds}^2}{2} \right)$$

This equation is called the **Sah Equation** after C.T. Sah, who first described the MOS transistor operation this way back in 1964. It is very important because it describes the basic behavior of the MOS transistor.

Note that for small values of V_{ds} , [a previous equation](#) and [\[link\]](#) will give us the same $I_d - V_{ds}$ behavior, because we can ignore the V_{ds}^2 term in [\[link\]](#). This is called the **linear regime** because we have a straight-line relationship between the drain current and the drain-source voltage. As V_{ds} starts to get larger however, the squared term will begin to kick in and the plot will start to curve over. Obviously, something is causing the current to drop off as V_{ds} gets larger. This is because the voltage difference between the gate and the channel is becoming less, which means there is less charge in the channel to provide conduction. We can graphically show this by making the channel layer look thinner as we move from the source to the drain. [\[link\]](#), and in fact, [\[link\]](#) would make us think that if V_{ds} gets large enough, that the drain current I_d should actually start decreasing again, and maybe even become negative!. This does not seem very intuitive, so let's take a look in more detail at the place where I_d becomes a maximum. We can define V_{dsat} as the source-drain voltage where I_d becomes a maximum. We can find this by taking the derivative of I_d with respect to V_{ds} and setting the derivative to 0.

Equation:

$$\begin{aligned} \frac{d}{dV_{ds}}(I_d) &= 0 \\ &= \frac{\mu_s c_{ox} W}{L} (V_{gs} - V_T - V_{dsat}) \end{aligned}$$

On dropping constants:

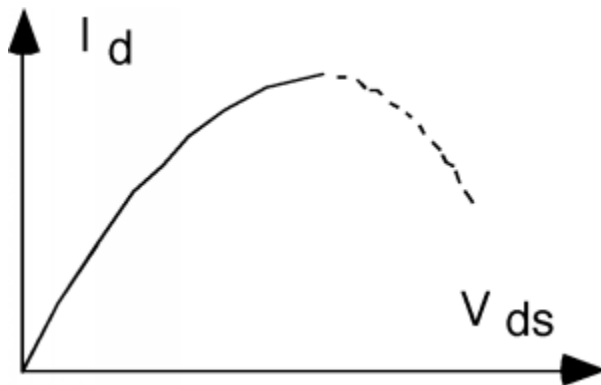
Equation:

$$V_{dsat} = V_{gs} - V_T$$

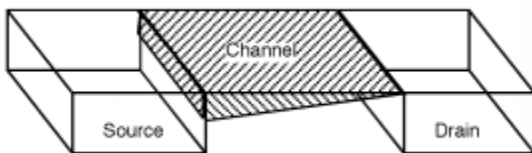
Rearranging this equation gives us a little more insight into what is going on.

Equation:

$$\begin{aligned} V_{gs} - V_{dsat} &= V_T \\ &= V_{gc}(L) \end{aligned}$$



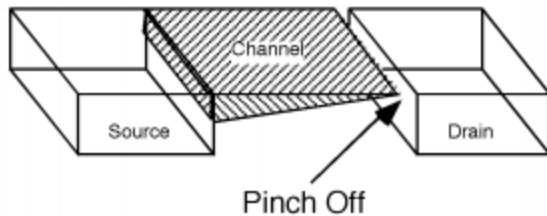
I-V characteristics showing turn-over



Effect of V_{ds} on the channel

At the drain end of the channel, when V_{ds} just equals V_{dsat} , the difference between the gate voltage and the channel voltage, $V_{gc}(L)$ is just equal to V_T , the threshold voltage. Any further increase in V_{ds} and the difference between the gate and the channel (**in the channel region just near the drain**) will drop below the threshold voltage. This means that when V_{ds}

gets bigger than V_{dsat} , the channel just near the drain region disappears! We no longer have sufficient voltage between the gate and the channel region to maintain an inversion layer, so we simply revert to a depletion condition. This is called **pinch off**, as seen in [\[link\]](#).

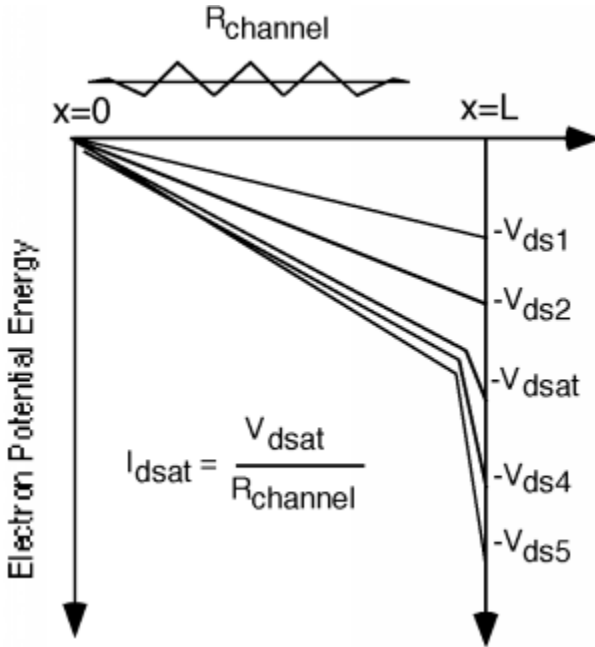


Channel in pinch-off

What happens to the drain current when we hit pinch off? It looks like it might go to zero, but that is not the right answer! Although there is no active channel in the pinch-off region, there is still silicon - it just happens to be depleted of all free carriers. There is an electric field, going from the drain to the channel, and any electrons which move along the channel to the pinch-off region are sucked across by the field, and enter the drain. This is just like the current that flows in the reverse saturation condition of a diode. There are no free carriers in the depletion region of the diode, yet I_{sat} **does** flow across the junction region.

Under pinch-off conditions, further increases in V_{ds} , does not result in more drain current. You can think of the pinched-off channel as a resistor, with a voltage of V_{dsat} across it. When V_{ds} gets bigger than V_{dsat} , the excess voltage appears across the pinch-off region, and the voltage across the channel remains fixed at V_{dsat} . If the channel keeps the same charge, and has the same voltage across it, then the current through the channel (and into the drain) will remain fixed, at a value we will call I_{dsat} .

There is one other figure which sometimes helps in seeing what is going on. We will plot potential energy for an electron, as it traverses across the channel. Since the source is at zero potential and the drain is at V_{ds} , an electron will loose potential energy as it flows from the source to the drain. [\[link\]](#) shows some examples for various values of V_{ds} :



Electron potential energy drop
going from source to drain.

For the first two drain voltages, V_{ds1} and V_{ds2} , we are below pinch-off, and so the voltage drop across R_{channel} increases as V_{ds} increases, and hence, so does I_d . At V_{dsat} , we have just reached pinch-off, and we are starting to see the "high field" depletion region begin to develop. Since electric field is just the derivative of the potential, the slope of curves in [\[link\]](#) gives you an idea of how big the electric field will be. For further increases in V_{ds} , V_{ds4} and V_{ds5} all of the additional voltage just shows up as a high field drop at the end of the channel. The voltage drop across the conducting part of the channel stays fixed (more or less) at V_{dsat} and so the drain current stays more or less fixed at I_{dsat} .

Substituting the expression for V_{dsat} into the expression for I_d we can get an expression for I_{dsat}

Equation:

$$I_{\text{dsat}} = \frac{\mu_s c_{\text{ox}} W}{2L} (V_{\text{gs}} - V_T)^2$$

We can define a new constant, k , where

Equation:

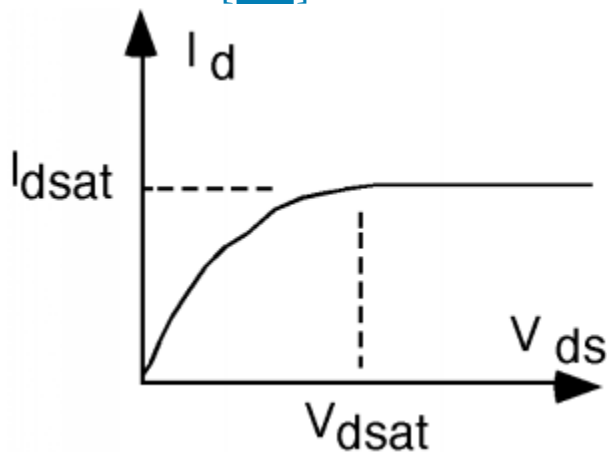
$$k = \frac{\mu_s c_{\text{ox}} W}{L}$$

So that

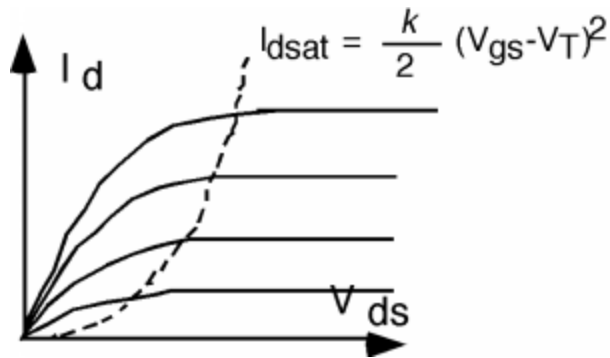
Equation:

$$I_{\text{dsat}} = \frac{k}{2} (V_{\text{gs}} - V_T)^2$$

What this means for [\[link\]](#) is that when V_{ds} gets to V_{dsat} , we simply hold I_d fixed from then on, with a value of I_{dsat} . For different values of V_g , the gate voltage, we are going to have a different $I_d - V_{\text{ds}}$ curve, and so once again, we end up with a family of "characteristic curves" for the MOSFET. These are shown in [\[link\]](#).



Complete I-V curve for the
MOSFET



Characteristic curves for a
MOSFET

This also gives us a fairly easy way in which to "sketch" a set of characteristic curves for a given device. Suppose we have a MOS field effect transistor which has a threshold voltage of 2 volts, a width of 10 microns, and a channel length of 1 micron, an oxide thickness of 150 angstroms, and a surface mobility of $400 \frac{cm^2}{V \cdot sec}$. using $\epsilon_{ox} = 3.3 \times 10^{-13} \frac{F}{cm}$, we get a value of $2.2 \times 10^{-7} \frac{F}{cm}$ for c_{ox} . This then makes k have a value of

Equation:

$$\begin{aligned}
 k &= \frac{\mu_s c_{ox} W}{L} \\
 &= \frac{400 \times 2.2 \times 10^{-7} 10}{1} \\
 &= 8.8 \times 10^{-4} \frac{amp}{volt^2}
 \end{aligned}$$

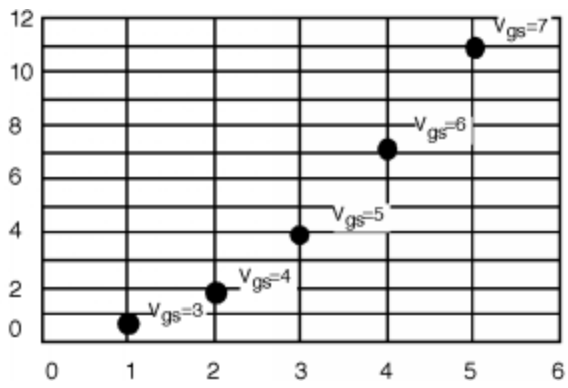
Plotting MOS I-V

Now we use two of the equations ([\[link\]](#) and [\[link\]](#)) that we found in the discussion about [MOS Regimes](#) to calculate a set of V_{dsat} and I_{dsat} values for various value of V_{gs} . (Note that V_{gs} must be greater than V_T for the two equations to be valid.) When we get the numbers, we build a little table.

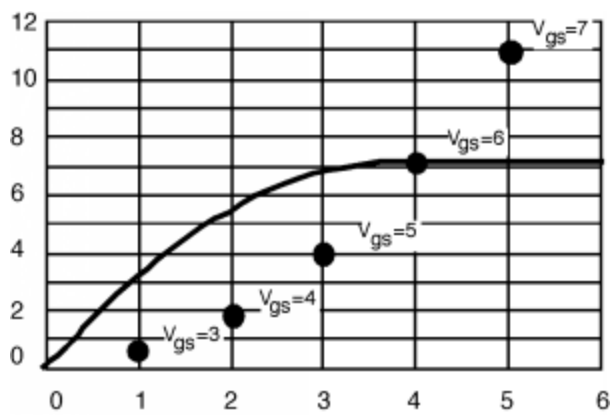
Once we have [the numbers](#), then we sketch a piece of graph paper with the proper scale, and [plot the points](#) on it. Once the I_{dsat} , V_{dsat} points have been determined, it is easy to sketch in the I-V behavior. You just draw a curve from the origin up to any given point, having it "peak out" just at the dot, and then draw a straight line at I_{dsat} to finish things off. One such curve is shown in [\[link\]](#). And then finally in [\[link\]](#) they are all sketched in. Your curves probably won't be exactly right but they will be good enough for a lot of applications.

V_{gs}	$V_{\text{dsat}}(V)$	$V_{\text{dsat}}(\text{mA})$
3	1	0.44
4	2	1.76
5	3	3.96
6	4	7.04
7	5	11

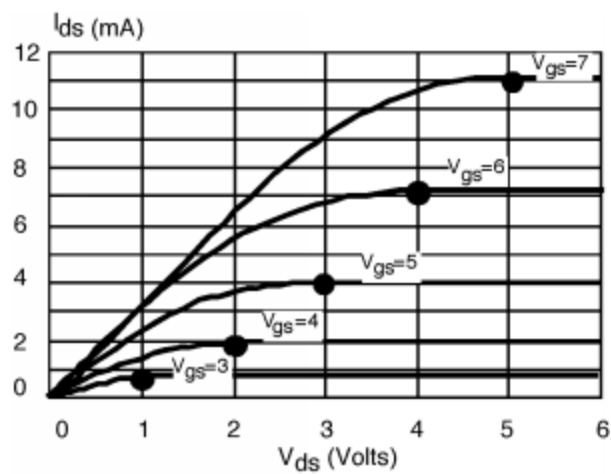
Results of calculating V_{dsat} and I_{dsat} .



Plotting I_{dsat} and V_{dsat} .

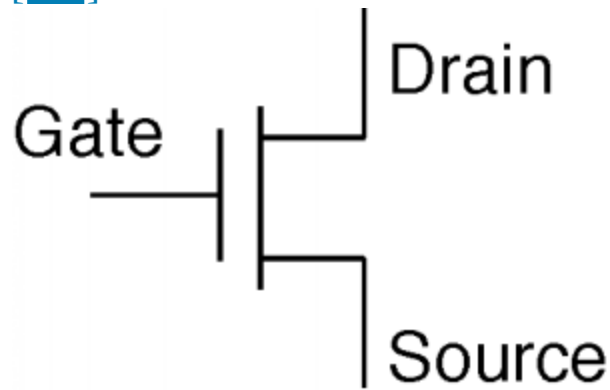


Sketching in one of the I-V curves.

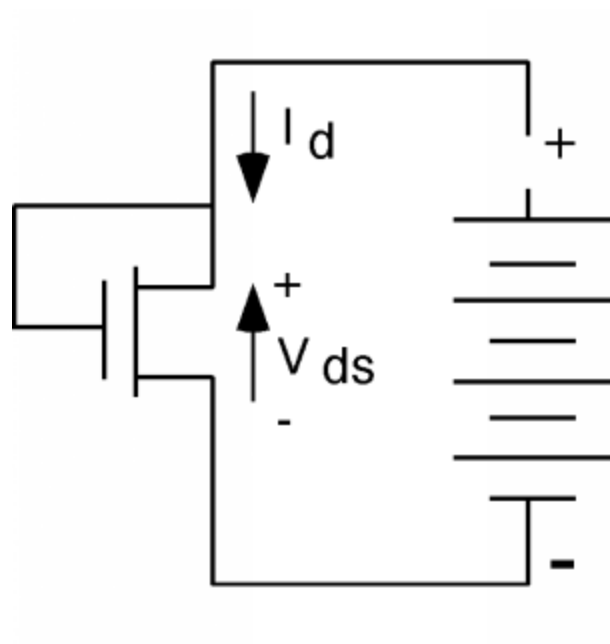


The complete set of curves.

There is a particularly easy way to measure k and V_T for a MOSFET. Let's first introduce the schematic symbol for the MOSFET, it looks like [\[link\]](#). Let's take a MOSFET and hook it up as shown in [\[link\]](#).



Schematic symbol for a MOSFET



Circuit for finding V_T and k

Since the gate of this transistor is connected to the drain, there is no doubt that $V_{gs} - V_{ds}$ is less than V_T . In fact, since $V_{gs} = V_{ds}$, their difference, is zero. Thus, for any value of V_{ds} , this transistor is operating in its saturated condition. Since $V_{gs} = V_{ds}$, we can rewrite [a previous equation](#) derived equation from the section on [MOS transistor](#) as

Equation:

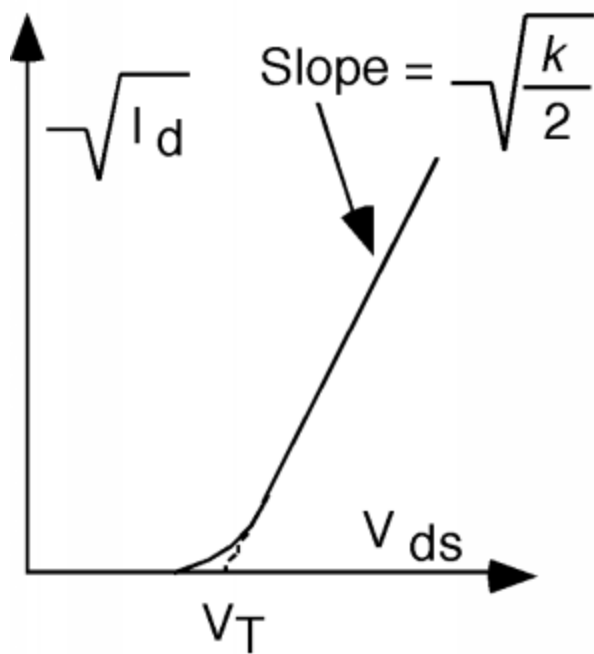
$$I_d = \frac{k}{2} (V_{ds} - V_T)^2$$

Now let's take the square root of both sides:

Equation:

$$\sqrt{I_d} = \sqrt{\frac{k}{2}} (V_{ds} - V_T)$$

So if we make a [plot](#) of $\sqrt{I_d}$ as a function of V_{ds} , we should get a straight line, with a slope of $\sqrt{\frac{k}{2}}$ and an x-intercept of V_T .



Obtaining V_T and k

Because of the expected non-ideality, the curve does not go all the way to V_T , but deviates a bit near the bottom. A simple linear extrapolation of the straight part of the plot however, yields an unambiguous value for the threshold voltage V_T .

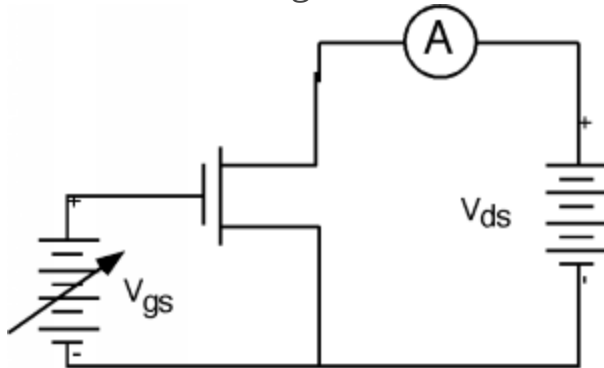
Models

A second, and some people think more accurate, way to find V_T is to look at the characteristics of the MOS transistor in its **linear regime**. The test circuit looks like what you see in [\[link\]](#). In this case, V_{ds} is kept quite small (0.2 Volts or so) and the gate voltage V_{gs} is swept over some range. If you look back at [equation in another module](#), which we can slightly re-write we see that

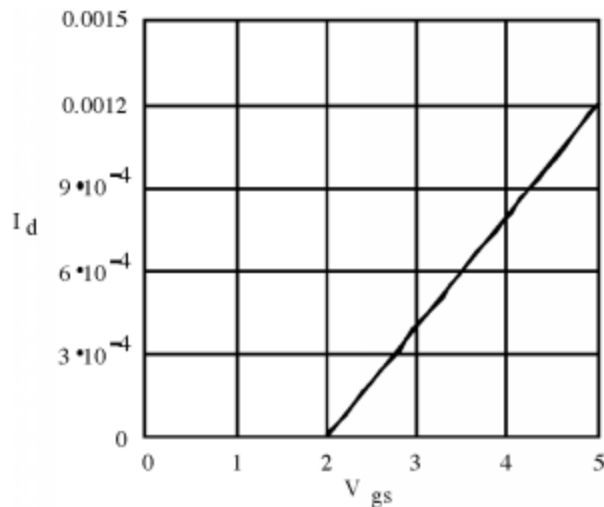
Equation:

$$I_d = \frac{\mu_s c_{ox} W V_{ds}}{L} (V_{gs} - V_T)$$

This equation will obviously give us a linear plot of I_d as a function of V_{gs} , which will look something like [\[link\]](#). Obviously, this is a device with a threshold voltage of about 2 volts. Can you figure out what k is for this transistor? If not, go back and re-read some stuff.

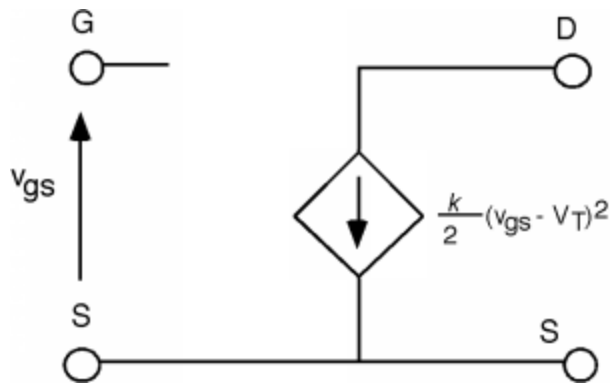


Circuit for finding V_T



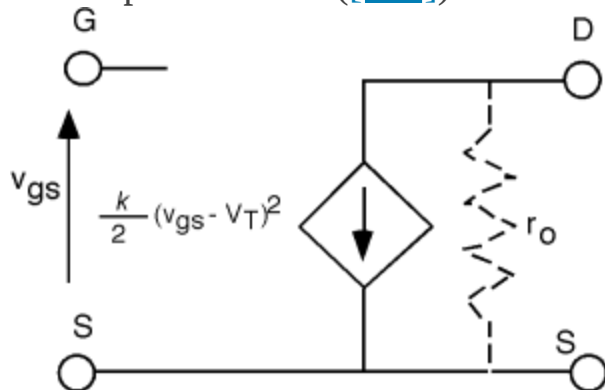
I_d as a function of V_{gs} for a MOS transistor in its linear range.

Now let's address a fundamental question concerning all of this: So What? What do we have here? One answer is that we have another device which in some way looks like the bipolar transistor we studied in the last chapter. In the saturation regime, the device looks and acts like a current source, and could probably be used as an amplifier. It is pretty easy to make a small signal model. The drain acts like a current source, which is controlled by V_{gs} . What should we do about the gate terminal? The gate really is not connected to anything inside the transistor, so it looks just like an open circuit. (In fact, there is a capacitance $C_{gate} = c_{ox}A_{gate}$, where $A_{gate} = WL$, the area of the gate, but in most low frequency linear applications, this capacitance is not significant.) Thus our small signal model for the MOSFET, if it is operating in its saturation mode, is as seen in [\[link\]](#).



Small signal MOSFET model

This seems to be a pretty good amplifier. It has infinite input impedance (and hence will not load down the previous stage of the amplifier) and it has a nice (but non-linear) voltage controlled current source for its output. [A figure](#) in the section on MOS regimes shows that as V_{ds} is increased, the channel length **does**, in fact, get a bit shorter. The increased V_{ds} makes the **pinch off** region expand a bit, which, of course, robs from the channel region. A shorter channel means slightly less channel resistance, and so I_d **actually** increases a bit with increasing V_{ds} instead of staying constant. We saw from the bipolar transistor, that when this occurs, we must add a resistor in parallel with our current source. Thus, let's complete the model with an additional r_o but in fact, we will put it in with a dashed line, because except for very short channel devices, it has very little effect on device performance ([\[link\]](#)).



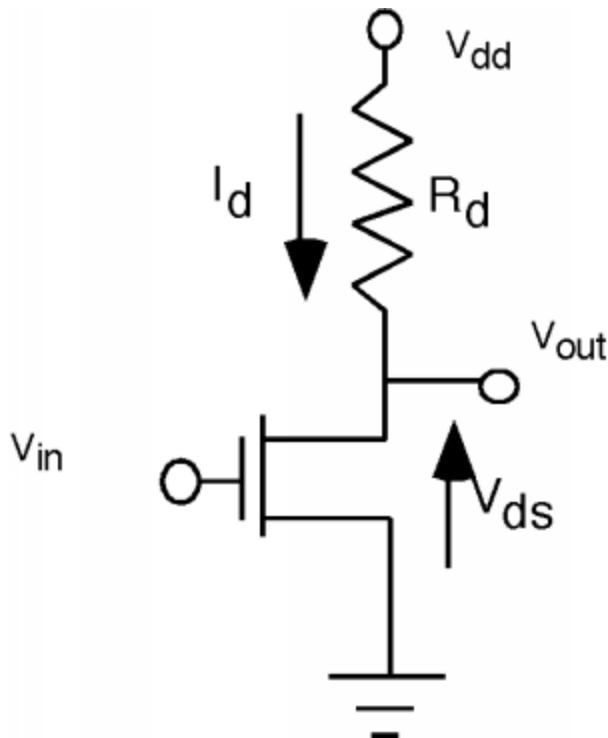
Adding an r_o

The MOSFET has several advantages over the bipolar transistor. One of the main ones, as we shall see, is that it is much easier to make. You only need two n-regions in a single p-type substrate. It is basically a surface device. This means you do not have to pile up different layers of n and p type material as you do with the bipolar transistor. Finally, we shall see that a variation on the MOSFET technology offers a **huge** advantage over bipolar devices when it comes to building logic circuits with a large number of gates (VLSI and ULSI circuits).

To see why this is so, we have to digress for just a little bit, and discuss [logic circuits](#).

Inverters and Logic

As you already know, or will find out shortly, from taking a class in digital logic, logic circuits are primarily based upon a circuit called an inverter. An inverter simply takes a signal and gives you the opposite one. For instance, if a high voltage (a "one") is placed on the input of an inverter, it returns a low voltage (a "zero"). [\[link\]](#) is a simple inverter based on a MOSFET transistor:



Inverter circuit

If V_{in} is zero, the MOSFET is turned off (V_{gs} is $< (V_T)$) and so no current flows through the resistor, and $V_{out} = V_{dd}$, a high. If V_{in} is high (and we assume that V_T for the MOSFET is significantly less than V_{in}) then the transistor is turned on, and if R and $\frac{W}{L}$ are chosen so that enough current flows through R to drop most of V_{dd} across it, then V_{out} will be low.

The way this is usually described is through a **transfer function** which tells us what the output voltage is as a function of the input voltage. Let's digress

for just a minute and see how such a function can be arrived at. Looking back at [\[link\]](#) it should be easy to see that

Equation:

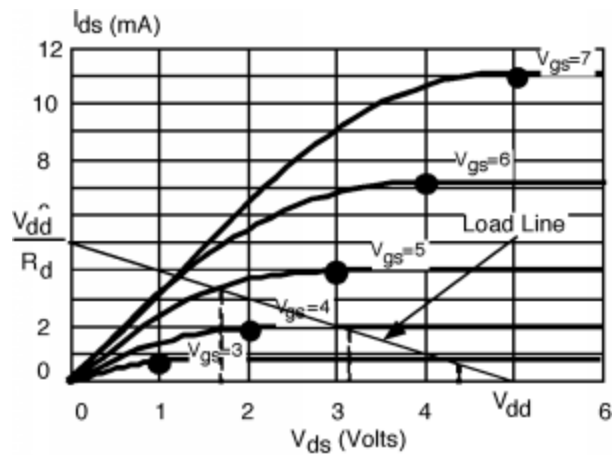
$$V_{dd} = I_d R_d + V_{ds}$$

We can re-write this as an equation for I_d .

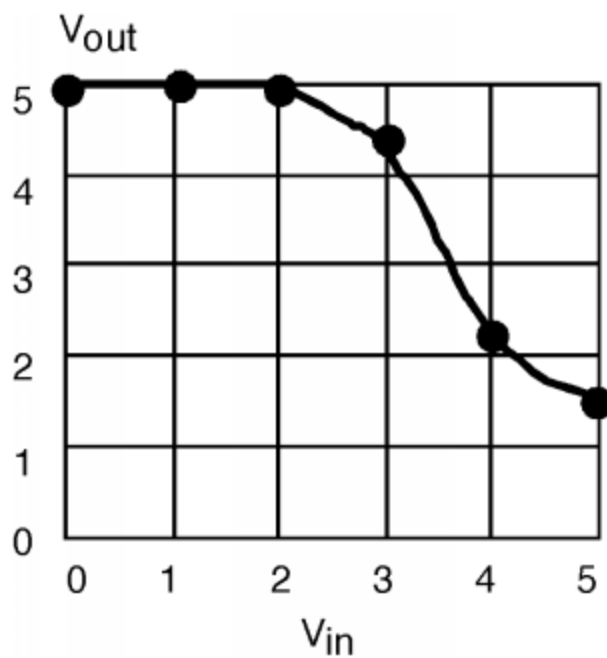
Equation:

$$I_d = \frac{V_{dd}}{R_d} - \frac{V_{ds}}{R_d}$$

This is called a **load-line** equation. It says that I_d varies linearly with V_{ds} (with a negative slope) and has a vertical off-set of $\frac{V_{dd}}{R_d}$. Let's suppose we have the MOSFET transistor for which we have already plotted the characteristic curves in a [previous plot](#). We will let $V_{dd} = 5$ Volts, and let $R_d = 1\text{k}\Omega$. From [\[link\]](#) we can see that when $V_{ds} = 0$, I_d will be 5 mA, and when $V_{ds} = V_{dd}$, I_d will be 0. This then gives us a straight line on the characteristic curve plot which is called the **load line**. This is shown in [\[link\]](#). By looking back at the schematic for the inverter in [\[link\]](#) we see that the same current I_d flows through the load resistor, R_d , and through the transistor. Thus, the correct value of current and voltage for the circuit for any given gate voltage is the simultaneous solution of the load line equation and the transistor behavior, which, of course, is just the intersection of the load line with the appropriate characteristic curve. Thus it is a simple matter of drawing vertical lines down from each V_{in} curve or V_{gs} value down to the horizontal axis to find out what the appropriate V_{dd} or output voltage will be for the inverter. Assuming that V_{in} only goes up to 5 Volts, the resulting curve that we get look like [\[link\]](#). This is not a great transfer characteristic. V_{in} has to get fairly large before V_{out} starts to fall, and even with the full 5 Volt input, V_{out} is still greater than 1 Volt. Picking a transistor with a small V_T and a bigger load resistor would give us a better response, but at least with this example you can see what is going on.



Characteristic curves with load line



Transfer characteristics for the inverter circuit.

Based on this simple inverter circuit, we can build circuits which perform the NOR and NAND function.

Equation:

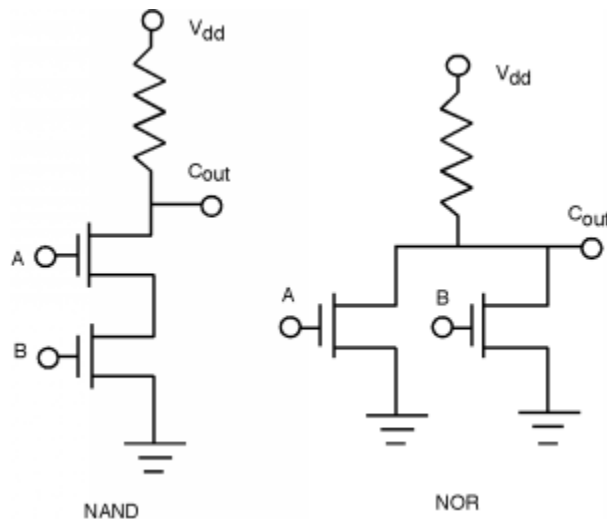
$$C_{\text{out}} = \neg(A + B)$$

and

Equation:

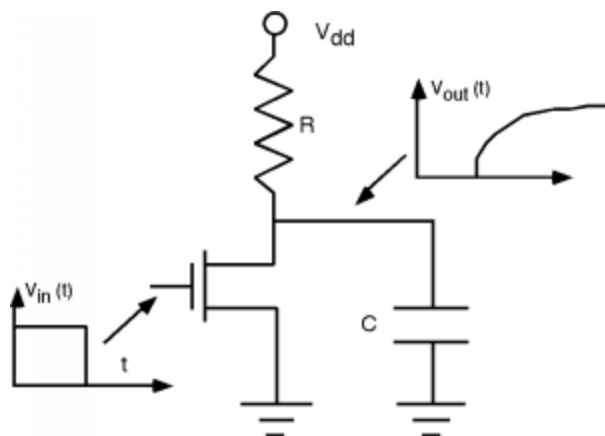
$$C_{\text{out}} = \neg(AB)$$

It should, by now, be obvious to you how the two circuits in [\[link\]](#) can perform the NAND and NOR function. It turns out that with the capability to do NAND and NOR, we can build up any kind of logic function we desire.



NAND and NOR circuits

Let's look at the inverter a little more [closely](#). Usually, the load for the inverter will be the next stage of logic which, along with the associated interconnect wiring, we can model as a simple capacitor. The value of the capacitance will vary, but it will be on the order of 10^{-12}F .



Driving a capacitive load

When the input to the inverter switches instantaneously to a low value, current will stop flowing through the transistor, and instead will start to charge up the load capacitance. The output voltage will follow the usual RC charging curve with a time constant given just by the product of R times C . If C is 10^{-13}F , then to get a rise time of 1 ns we would have to make R about $10^4\Omega$.

As we shall see later, it is virtually impossible to make a 10 k Ω resistor using integrated circuit techniques. Remember:

Equation:

$$R = \frac{\rho L}{A}$$

And thus, to get a really big resistance we need either a very tiny A (Too hard to achieve and control.), a really BIG L (Takes up too much room on the chip) or a huge ρ (Again, very hard to control when you get to the very low doping densities that would be required.)

Even if we could find a way to build such big integrated circuit resistors, there would still be a problem. The current flowing through the resistor when the MOSFET is on would be approximately

Equation:

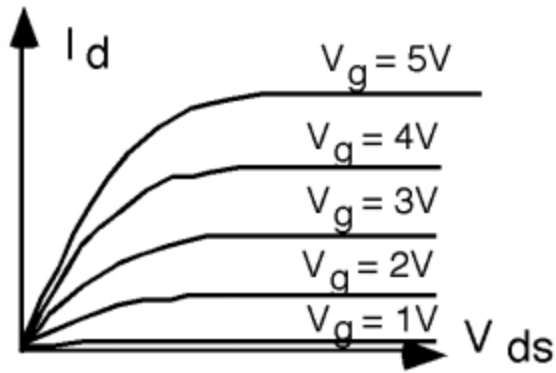
$$\begin{aligned} I &= \frac{V}{R} \\ &= \frac{5V}{10^4\Omega} \\ &= 5 \times 10^{-4} A \end{aligned}$$

Which doesn't seem like much current until you consider that a Pentium© microprocessor has about 6 million gates in it. This would mean a net current of -300 Amps flowing into the CPU chip! We've got to come up with a [better solution](#).

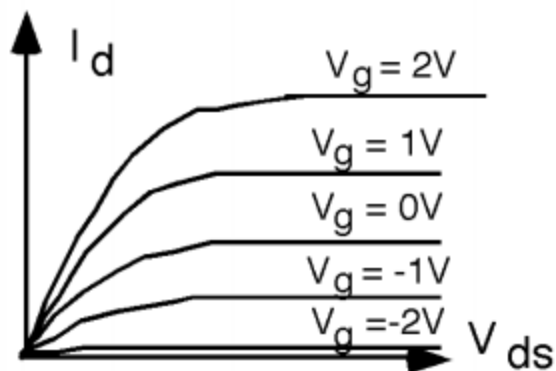
Transistor Loads for Inverters

There are other kinds of MOSFET's besides the one we have studied so far. Strictly speaking, what we have seen up to now is called an **n-channel enhancement mode MOSFET**. It turns out that you can build a MOSFET which looks just like a [previous figure](#), except that by putting some additional impurities under the gate region, we can arrange it so that there is a channel formed, even with $V_g = 0$. The transistor now has a **negative** V_T . The process by which the additional impurities are added is called a V_T **adjust**.

A MOSFET with a negative V_T can be expected to have $I_d - V_{ds}$ curves similar to those for a positive V_T device, except for one thing. For $V_{gs} = 0$, the device is already turned on, and so we get a usual MOSFET-type curve. **Positive** gate voltage turns it on even more, while negative V_{gs} tends to reduce the drain current. It takes a **negative** gate voltage to turn the thing off. [\[link\]](#) shows comparative characteristic curves for an enhancement and depletion mode devices.



Enhancement Mode MOSFET



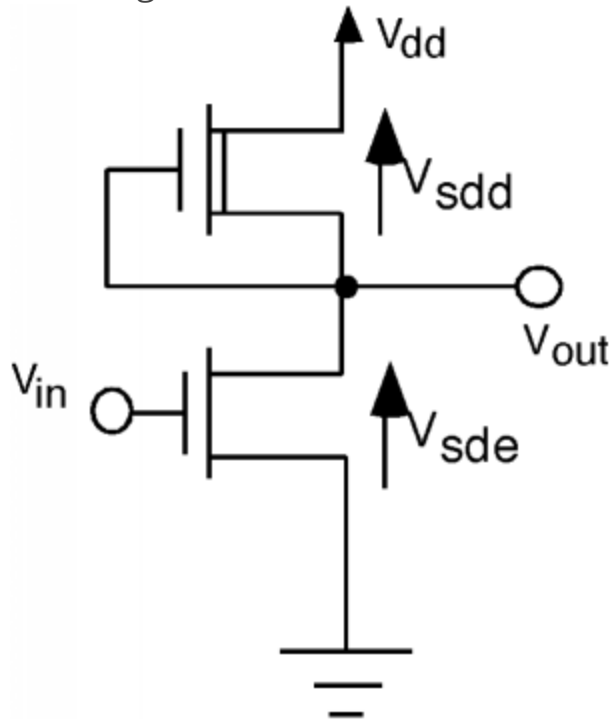
Depletion Mode MOSFET

Enhancement and depletion
characteristic curves

For an enhancement mode transistor, you have to get $V_g > V_T$ (-1 Volt in this example) to **enhance** the conductivity or channel to make it conduct. For a depletion mode device, a gate voltage V_{gs} of 0, still finds the device conducting. You have to put some negative voltage on the gate to **deplete** the channel, in order to turn it off. We now have a **depletion mode n-channel MOSFET**.

How would we use a depletion mode device in an inverter gate? The answer is fairly straight-forward. In the schematic in [\[link\]](#), we indicate a depletion mode MOSFET by adding a second line, under the gate, to suggest that a channel already exists in the device, even with no V_g . Note that the gate of

the depletion mode transistor (also sometimes called the **pull up** transistor) is connected to its source, so, in fact, V_{gs} does equal 0 for this device. The input transistor (or the **pull down** transistor) is just an enhancement mode MOSFET like we had before. It is not hard to choose appropriate W and L so that I_{dsat} for the pull up transistor is on the order of the 500 μA that we need to get our 1 ns rise time on the capacitive load.



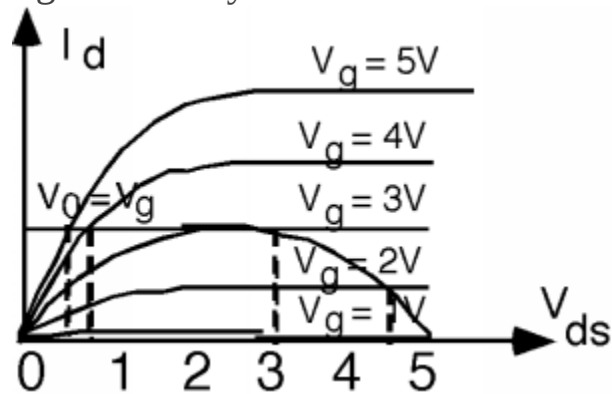
Depletion mode load

In order to get the transfer characteristic for this circuit, we first note that **Equation:**

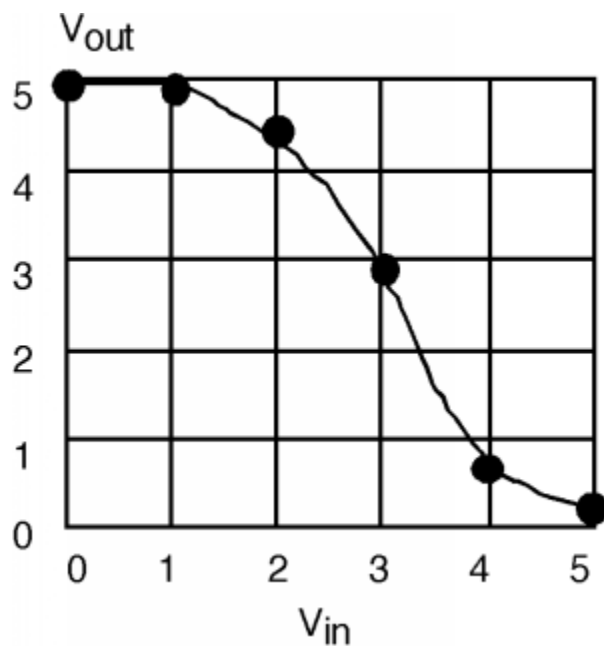
$$V_{sdd} = V_{dd} - V_{sde}$$

where V_{sde} is the source-drain voltage for the pull-down, or enhancement transistor, and V_{sdd} , is the source-drain voltage for the depletion-mode transistor. If we want to plot the **load-line** for the pull-down transistor that is created by the pull-up or depletion mode transistor, we should take its $V_{gs} = 0$ characteristic curve, shift it over by an amount V_{dd} , and then

reverse its polarity. When we do this we get the following shown in [\[link\]](#). Noting the intersection points of the **load line** and the characteristic curves allows us the opportunity for drawing the [transfer characteristic](#). This is a better looking curve. It is symmetric around the mid-voltage point, and gets closer to zero for its output "low" condition. The transition from "high" to "low" is also somewhat more abrupt, which is advantageous. Can you figure out why?



Characteristic curve and load line for a depletion MOS load



Transfer characteristics for a

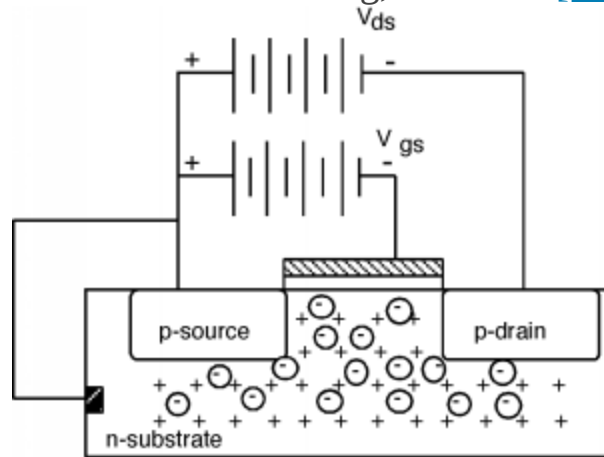
depletion load inverter

Well, we solved one problem. At least we have a pull up structure that we can manufacture. It turns out not to be too hard to build an enhancement MOSFET that has an equivalent resistance in the range we need without taking up too much chip area. We have not solved the other problem however. We are still looking at a **huge** current draw for large circuits. Since on average, half of the inverter gates will be "on" in a logic circuit, we still have a large current sink to ground. This is something that would be completely prohibitive in a modern-day VLSI integrated circuit.

Fortunately, we have not run out of options for [MOS structures](#) yet.

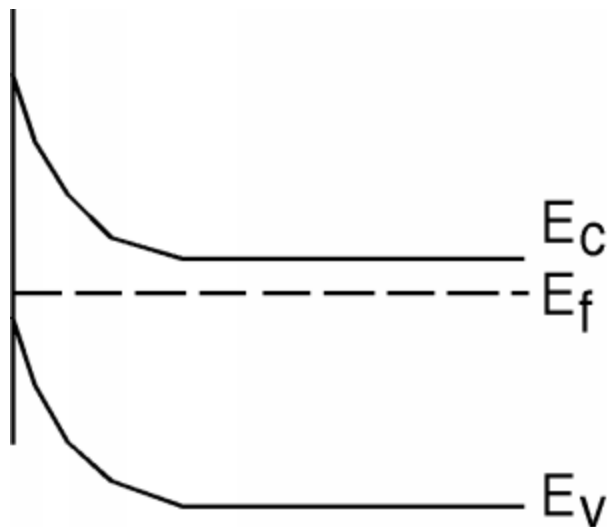
CMOS Logic

Consider the following, shown in [\[link\]](#).



A PMOS transistor

This looks a lot like our previous MOSFET except that now we have an n-type substrate and the source and drain regions are p-type. If we apply a **negative** V_{gs} (with the source connected to the n-type substrate) then the induced negative charge on the gate will drive away the electrons, and if the bands under the gate are bent up sufficiently, form an [inversion layer](#) of **holes** thus making an enhancement mode **p-channel MOSFET**, or a PMOS transistor. (As opposed to an NMOS transistor which we studied first.). Note that a PMOS transistor will have a negative V_T . That is, the gate voltage has to be **less than the source/substrate voltage** in order to turn the device on. The more negative V_{gs} , the more current we will have flowing through the device.



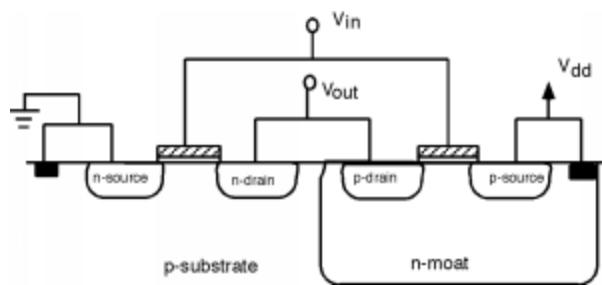
Inversion of an n-type layer

It turns out that a combination of both an n-channel and a p-channel device on the same circuit can be very advantageous. Such technology is called **CMOS**, for "complementary MOS". Here is how we use a p-channel transistor in the inverter circuit.

First of all, however, we have to see how we would make one. There is a fundamental problem in trying to use both n-channel and p-channel devices in the same circuit. What is it? It would seem we need two different kinds of substrates, both a p-type substrate for the n-channel transistor, and an n-type substrate for the p-channel device. There is a way around this problem by making what is called a **tank** or a **moat**. A moat is a relatively deep region of one type of material placed into a host substrate of the opposite type ([\[link\]](#)). We can put n-type source/drain regions into the p-substrate and p-type source/drain regions into the n-moat. In [\[link\]](#), we will also show the gates, and how the whole inverter is connected together.

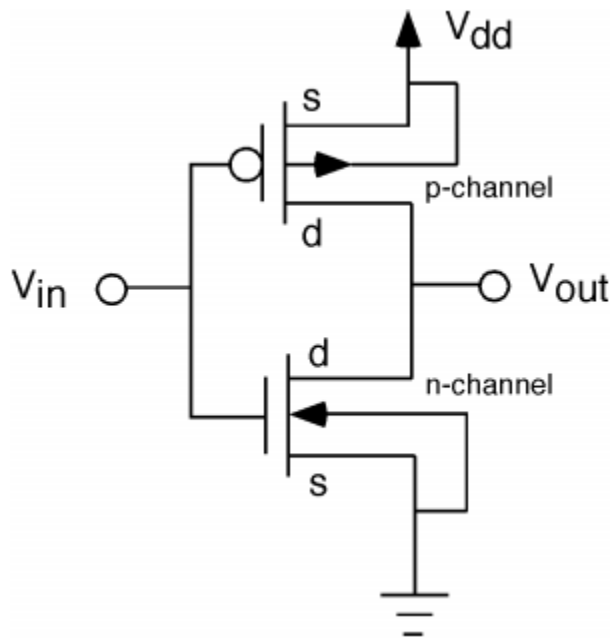


Preparing for a CMOS inverter



A CMOS inverter

Now let's draw the [schematic](#): A p-channel device is drawn just like an n-channel device, except we put a little "bubble" on the gate to signify that it is a MOSFET of a different color. Although we usually don't do this all the time, we have also shown the substrate connections in this diagram. These connections show that a MOSFET is at least a four terminal device, not a three terminal one as people often assume. Since, in a p-channel device, the substrate is n-type, we show the substrate connection as an outward pointing arrow. The p-type substrate for the n-channel device is shown as an inward pointing arrow. The n-channel substrate is connected to ground, the p-channel substrate is connected to V_{dd} . Note that since the n-moat is at V_{dd} and the p-substrate is at ground, the moat-substrate p-n junction is reverse biased, and so no current should flow between them.



Schematic of a CMOS inverter

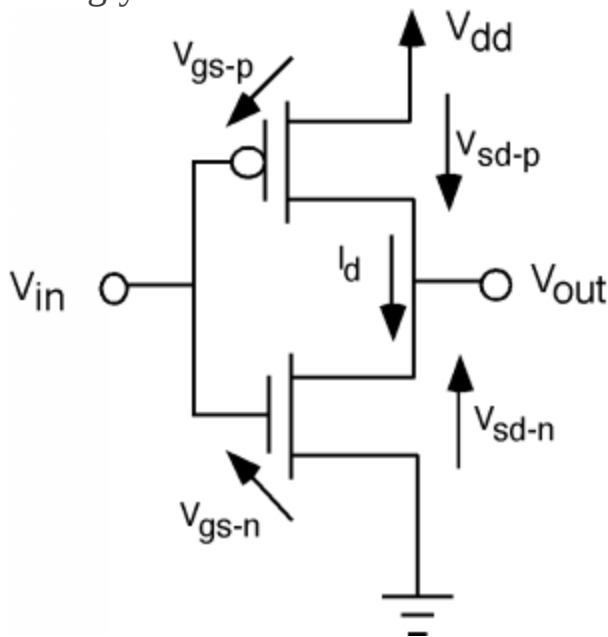
We usually do not label the source and drain either, but we do here, just for completeness. Note that unlike the bipolar transistor, the FET is truly a symmetric device. There is really no way to tell the source from the drain. By convention, we call the element which is connected to the substrate (or moat) the source, and the other the drain. You will sometimes hear the region under the gate (either substrate or moat) referred to as the **backbody**.

Now let's see how this circuit works. If V_{in} is high (at or near V_{dd}) the NMOS transistor will be turned on. The voltage between the gate and substrate of the p-channel device is at or near zero. The gate is at V_{dd} and so is the moat! Hence the upper transistor will be turned off. The output will thus be **low**.

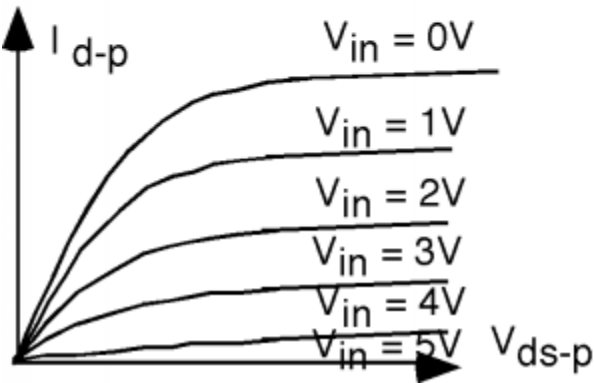
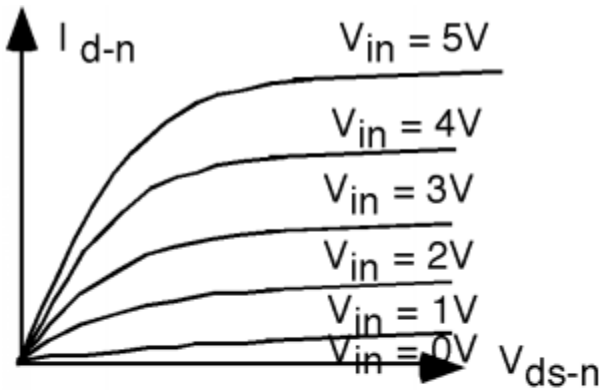
If the input voltage is at or near ground (a "low") then the n-channel device is turned off. The voltage between the gate and substrate of the p-channel device is now $\approx (-V_{dd})$. (The gate is $\approx (0)$ and the substrate is at V_{dd} .) If the PMOS transistor has a threshold voltage V_T of, say, -2 V, then it will be

turned **on** and the output will be **high**. Note however, that in either state, high or low, **there is no static current flowing through the inverter**.

The transfer characteristics for this circuit. Are a little more complicated. First, let's make sure we have our voltages and currents [defined](#). From the figure, V_{gs-n} the n-channel gate-source voltage is just V_{in} . V_{gs-p} the gate-source voltage for the p-channel device is $V_{in} - V_{dd}$. $I_{d-n} = I_{d-p} = I_d$. V_{ds-p} the drain source voltage for the p-channel transistor can be written as $V_{dd} - V_{out}$. $V_{ds-n} = V_{out} - 0$. We have two sets of [characteristic curves](#): Note that since $V_{gs-p} = V_{in} - V_{dd}$, when $V_{in} = 0V$, $V_{gs-p} = -5V$ and so the transistor is strongly turned on.

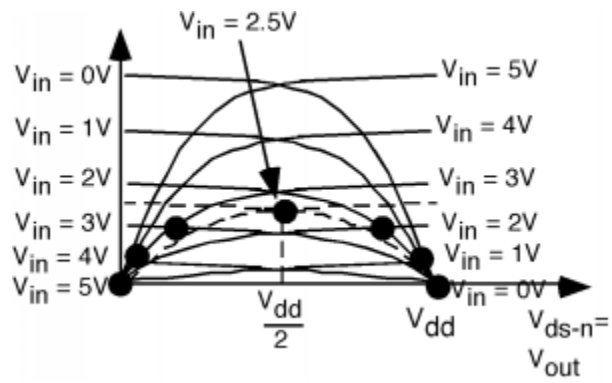


Defining voltages

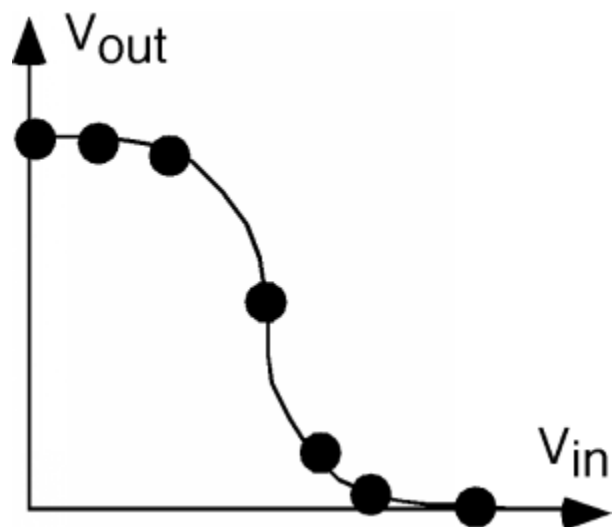


Drain currents for the two transistor as a function of input voltage and V_{ds}

We have a number of different "load lines" in this case, because for each V_{in} we have a different curve for both the n and p channel transistors. This is shown in [\[link\]](#). The black spots show the point of intersection. Follow a few of the curves along to see if you agree with where the spots have been placed. We have also added a pair of dotted curves for $V_{in} = 2.5V$ so we can get the "turn-over" point. Projecting the location of the black dots to the V_{ds-n} (or V_{out}) axis will give us a value for V_{out} for each of the input voltages, V_{in} . The resulting curve is shown in [\[link\]](#). This gives us a good "feel" for how the inverter works, and how the output varies with the input. Note that this transfer curve is quite symmetric about 2.5 volts, and goes all the way from +5 to 0 volts on the output.



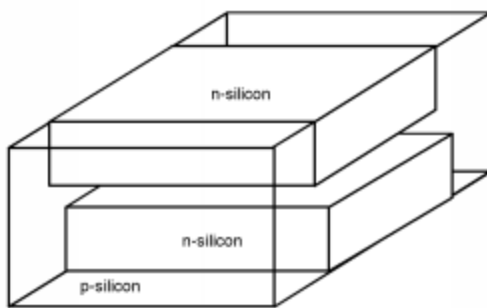
Getting the transfer function



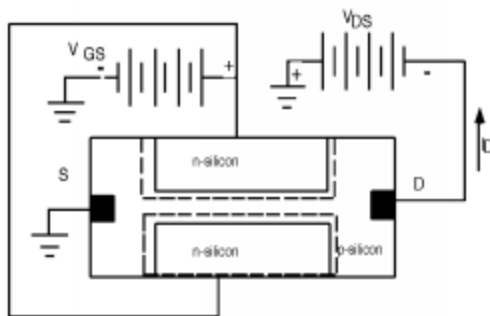
CMOS inverter transfer
characteristics

JFET

There is a lot more that we could do with field effect devices, but it is probably time to move on to new topics. For one final point however, we might just look at something called the JFET, or junction field effect transistor. The JFET structure looks like [\[link\]](#). It consists of a piece of p-type silicon, into which two n-type regions have been diffused. However, instead of being both on the same surface, as with a MOSFET, the two regions are opposite one another on either side of the crystal. In cross-section, the JFET looks like [\[link\]](#). We also show the biasing here.



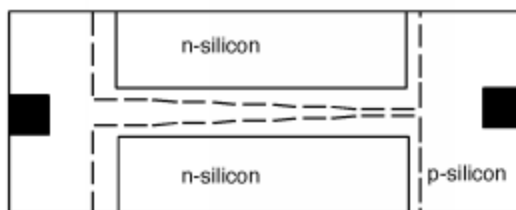
JFET



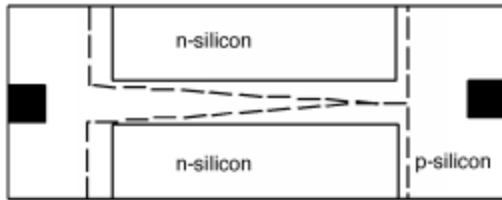
Biasing a JFET

The two n-regions are connected together, and are reverse biased with respect to the p-type substrate. A second battery, V_{ds} is used to pull current out of the source, by applying a negative voltage between the drain and the source. The reverse biased n-p junctions create a depletion region which extends into the p-type material through which the holes travel as they go from source to drain (a channel?). By adjusting the value of V_{gs} , one can make the depletion region smaller or larger, thus increasing or decreasing the drain current.

The observant student will also note that the polarity of the V_{ds} battery makes it so that there is more reverse bias across the p-n junctions at the drain end of the channel than at the source end. Thus, a more accurate depiction of the JFET would be what is shown in [\[link\]](#). When the drain/source voltage gets large enough, the two depletion regions will join together, and, just as with the MOSFET, the channel pinches off, as shown in [\[link\]](#).



Depletion region controls
current



Pinch-Off

Surprising as it may seem, when you work out the equations which describe how the depletion region extends with V_{gs} and how the pinch-off mechanism changes I_D , you end up with behavior, and equations, which are quite similar to those of a depletion-mode MOSFET.

Using JFETs is a little more cumbersome than a normal MOSFET. You must make sure that the gate-substrate junction always remains reverse biased, and since the JFET can only be a depletion-mode device, you have to have a voltage on the gate if you want to turn the transistor off. The JFET **does** have one advantage over the MOSFET however. A while back we calculated the value for C_{ox} the oxide capacitance and found that it was on the order of $10^{-7} \frac{F}{cm^2}$. A typical MOSFET gate might be $1 \mu m$ long by $20 \mu m$ wide, and so it would have a gate area of $20 \mu m^2$ or $2 \times 10^{-7} cm^2$. Thus, the total gate capacitance is only about $10^{-14} F$.

Electrostatic Discharge and Latch-Up

As you are probably aware, you have to be very careful when handling MOS circuits, to be sure that you are properly grounded, and that you do not transfer any static electricity to the chip. The **standard human-body model** assumes a static charge transfer of about 0.1 micro-Coulombs ($10^{-7}C$) upon static electricity discharge between a human and a chip. This does not seem like enough charge to do any harm until we remember the old formula:

Equation:

$$Q = CV$$

or

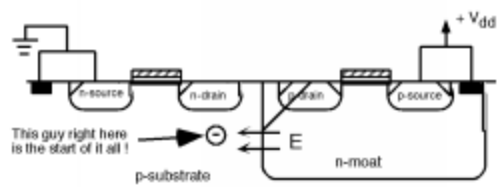
Equation:

$$V = \frac{Q}{C}$$

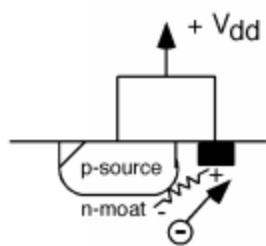
Last time I looked 10^{-7} divided by 10^{-14} is about 10^7 volts! Add to this the fact that the gate oxide thickness is only about 10^{-6} cm, so that we have electric fields in the gate oxide which are on the order of $10^{13} \frac{V}{cm}$! No wonder the things break. This problem is called **electrostatic discharge**, or ESD, and is one of the major concerns of IC manufacturers. Protecting against ESD is still very much a "black art" and is something that people are still studying quite a bit. JFET's are much more rugged structures, and have much higher gate capacitances, and are not nearly so prone to ESD failure.

Since we are on the subject of problems, let's take a look at one more "glitch" that plagues IC designers. We have to go back to the CMOS circuit. Remember, the moat/substrate junction is reverse biased, so we will have an electric field in the depletion region of that junction, pointing as shown in [\[link\]](#). Suppose, somehow, we have one or more stray electrons in the p-type substrate. They will be swept across the substrate/moat junction by the electric field, and be attracted to the moat contact by V_{dd} . Let's focus on what happens as the electron flows out the V_{DD} contact ([\[link\]](#)). As the

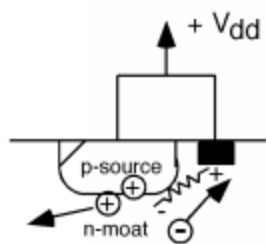
electron moves through the (resistive) n-type moat material, it develops a voltage drop between the n-type material under the source, and the V_{DD} contact (Which is also at the source potential since they are connected together by the interconnect on the surface of the wafer.) Electron flow in one direction means current flow in the other and so this makes the region under the source slightly negative with respect to the source region itself. This, of course, forward biases the source/moat junction slightly, which causes a hole or two to be injected into the moat from the p-source ([\[link\]](#)). The holes will be attracted by the field across the moat-substrate depletion layer, and, once they get there, they will be swept into the p-substrate ([\[link\]](#)). Once the holes get into the p-substrate, they will be attracted to the ground connection so that they can leave the semiconductor. As these holes flow past the n-source, and through the resistive p-substrate, they build up a potential between the ground contact ([\[link\]](#)), and the material under the source with a polarity which tends to forward bias the source-substrate junction, and cause electrons to be injected into the substrate. The electrons, in turn, are attracted to the field across the substrate-moat junction ([\[link\]](#)). Some of the electrons may recombine in the p-region, but in today's high-quality substrates, there are very few active recombination centers, and so even though the electrons are minority carriers, they have quite a long minority carrier lifetime, and most of them make it to the substrate-moat junction. and are swept into the moat. Once inside the n-moat, the electrons are then attracted to the $+V_{dd}$ contact, where, of course, they build up a bigger forward bias across the source-moat junction, causing more holes to be emitted from the source into the moat ([\[link\]](#)). These holes are swept across the moat-substrate junction, flow to the ground contact and, well ... you get the idea! It does not take long before we have a dead short circuit between V_{dd} and ground. This is not healthy for integrated circuit chips in the least, and is a process called **latch up** ([\[link\]](#)).



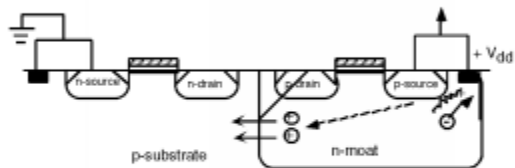
The start of trouble!



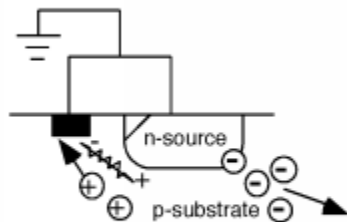
Electron flow builds up voltage



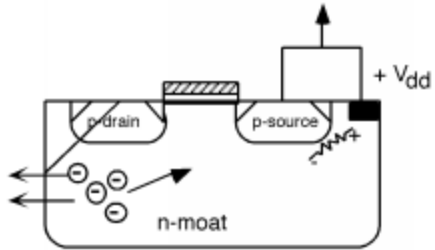
The forward biased source
injects some holes



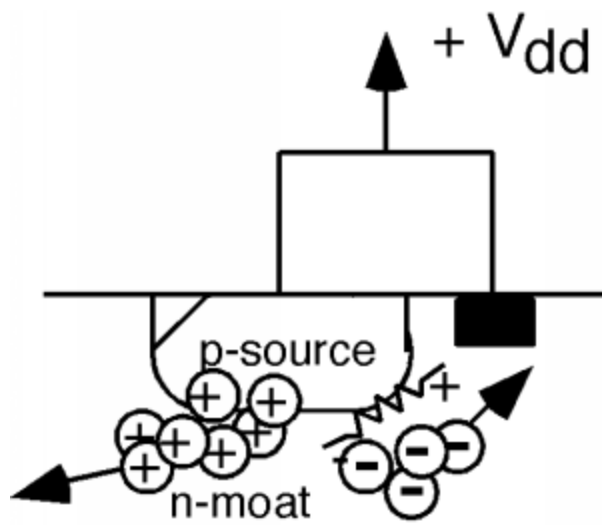
The holes are swept into the
substrate



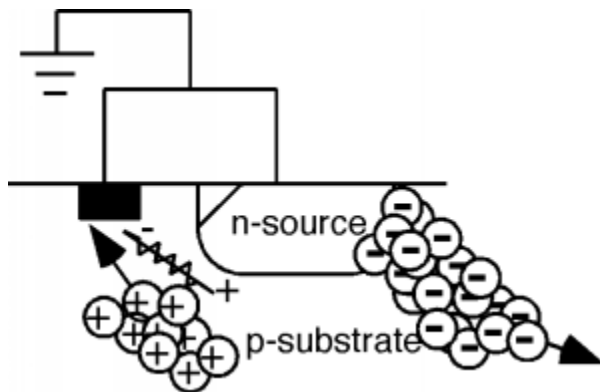
Voltage drop at the n-channel
source end.



The electrons are swept into the moat

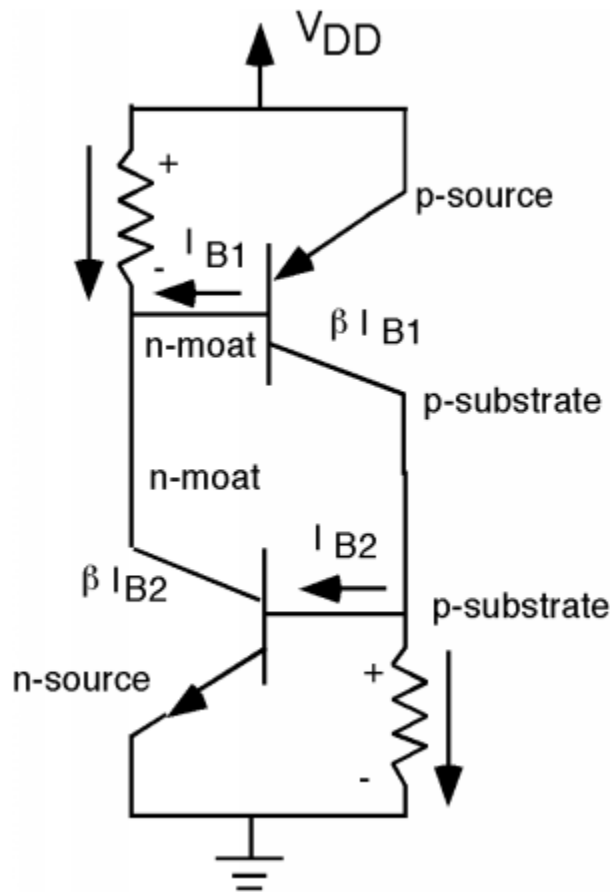


More current means a bigger voltage and more holes injected.



Latch Up!

There is an interesting circuit you can draw which shows what is happening from a somewhat different point of view. Note that we can consider the p-source, n-moat, and p-substrate as a pnp bipolar transistor. Also the n-source, p-substrate and n-moat also make a fine npn bipolar transistor. The two transistors are intermingled however, with the base of the pnp and the collectors of the npn sharing the same n-moat, and the collector of the pnp and the base of the npn sharing the same p-substrate. The n-moat and p-substrates are both collectors **and** bases at the same time. A little careful inspection of the cross section of the CMOS inverter will lead you to the following schematic shown in [\[link\]](#). We need something to get this circuit started, so say we have a little collector current coming out of the top pnp transistor. This current flows down, through the resistor to ground. As it flows through the resistor it builds up a little voltage which forward biases the base-emitter junction of the lower, npn, transistor, and causes some collector current to flow into it. This current comes from V_{dd} through the upper resistor, and builds up a voltage across that resistor which will forward bias the base-emitter junction of the top, pnp, transistor. This, in turn, causes some additional collector current to flow out of the pnp transistor, and away we go! Latch-up is bad, and is something which IC designers work very hard to avoid.



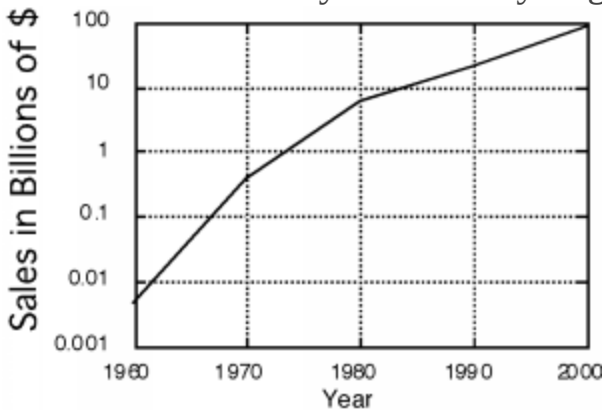
Schematic of latch up circuit

You might wonder what **actually** starts a circuit going into latch-up. Refer back to the [CMOS inverter](#), and note that the n-drain on the NMOS is connected to the output. The output **could** be a real output, going beyond the chip into the "real world". If the "customer" who is using the chip is careless, and somehow drags the output down below ground, the drain/p-substrate junction will be forward biased, electrons will be injected into the p-substrate, and we are back at [\[link\]](#). IC designers try to keep the n-moat/ V_{dd} contact as close to the PMOS source, and the p-substrate/ground contact as close to the NMOS source as they can to reduce the resistance between the contact and the source regions, and hence lower the chance of the circuit going into latch-up.

Introduction to IC Manufacturing Technology

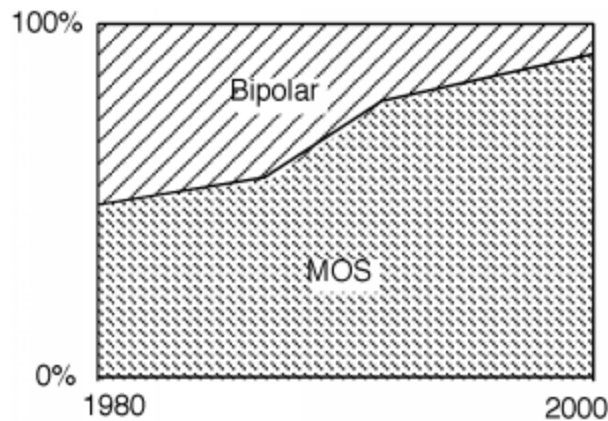
It would probably be interesting to spend a little time seeing how integrated circuits are made. This chapter will be long on description, and rather short on equations (yay!). This is not to say that there is not a lot of analytical work in the IC fabrication process. It's just that things get **very** complicated in a hurry, and so we probably are better off just looking at most processes from a qualitative point of view.

Let's start out by taking a look at the state of the industry, and remark on a few trends. [\[link\]](#) is a plot of IC sales in the United States over the past 30 years. This might not be a bad field to get into! Maybe there will be a job or two out there when you are ready to graduate.



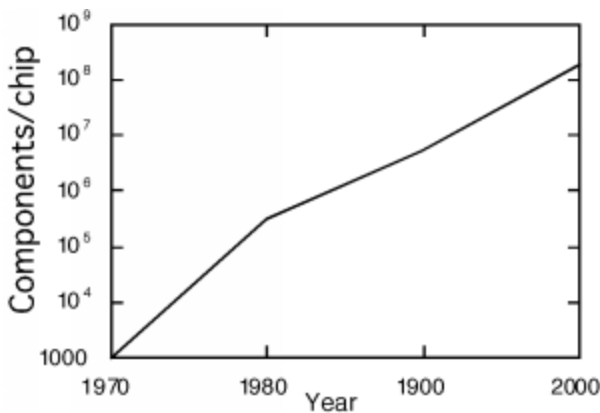
Growth of IC Business

There has been a steady shift away from bipolar technology to MOS as is shown in [\[link\]](#). Currently, about 90% of the market is composed of MOS devices, and only about 10% of bipolar. This is likely to be the case for some time to come. The change in slope, where MOS starts taking over from bipolar at a more rapid rate about 1987 is when CMOS technology really started to come into heavy use. At that point, bipolar TTL logic essentially faded to zero.



Percentage of Business

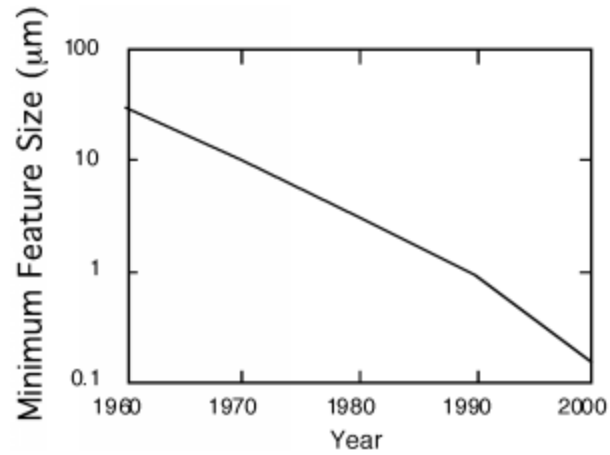
As you probably are aware, devices have been getting smaller and smaller, and chips have been getting bigger and bigger with time. A most [impressive plot](#) is one which shows the number of components/chip as a function of time.



Number of transistors/chip

One of the main drivers for this has been feature size, which shows the same nearly exponential behavior as components/chip. This is plotted in [\[link\]](#) for your education. What is interesting to note about this is that certain "doom sayers" have been predicting an abrupt halt to this curve for some time now. It stands to reason that you can not image something which

is finer than λ , the wavelength of the light you use to project it with. However, by going to the ultraviolet, and using a variety of image enhancing techniques, lithographic engineers continue to be able to make finer and finer structures.



Feature size with time

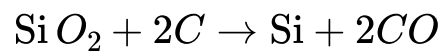
Silicon Growth

How is it possible for the IC industry to continue to make such gains, and how do they build so many circuits on one chip anyway? In order for us to be able to understand this, we have to take a look at the **monolithic fabrication process**. Lith comes from the Greek word for stone, and mono means one, of course. Thus, monolithic construction refers to building the circuit in "one stone" or in one single crystal substrate.

In order for us to do this however, we first of all need the "stone", so let's see where that comes from.

We start out with a natural form of silicon which is very abundant (and relatively pure); quartzite or Si O_2 (sand). In fact, silicon is one of the most abundant elements on the earth. This is reacted in a furnace with carbon (from coke and/or coal) to make what is known as **metallurgical grade silicon** (MGS) which is about 98% pure, via the reaction

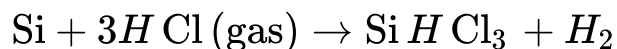
Equation:



We have seen that on the order of 10^{14} impurities will make major changes in the electrical behavior of a piece of silicon. Since there are about 5×10^{22} atoms/cm³ in a silicon crystal, this means we need a purity of better than 1 part in 10^8 or 99.999999% pure material. Thus we have a long way to go from the purity of the MGS if we want to make electronic devices that we can use in silicon.

The silicon is crushed and reacted with H Cl (gas) to make trichlorosilane, a high vapor pressure liquid that boils at 32°C as in:

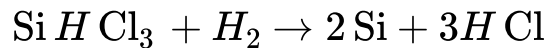
Equation:



Many of the impurities in the silicon (aluminum, iron, phosphorus, chromium, manganese, titanium, vanadium and carbon) also react with the

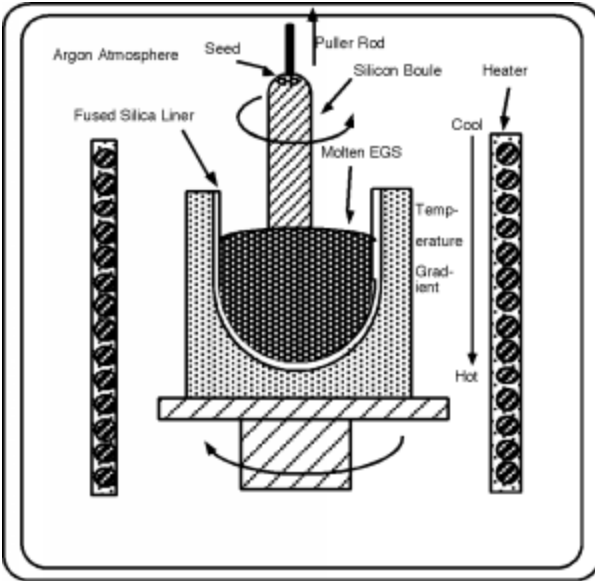
HCl , forming various chlorides. One of the nice things about the halogens is that they will react with almost anything. Each of these chlorides have different boiling points, and so, by fractional distillation, it is possible to separate out the SiHCl_3 from most of the impurities. The (pure) trichlorosilane is then reacted with hydrogen gas (again at an elevated temperature) to form pure **electronic grade silicon** (EGS).

Equation:



Although the EGS is relatively pure, it is in a polycrystalline form which is not suitable for device manufacture. The next step in the process is to grow single crystal silicon which is usually done via the **Czochralski**(pronounced "cha-krawl-ski") method to make what is sometimes called CZ silicon. The Czochralski process involves melting the EGS in a crucible, and then inserting a seed crystal on a rod called a puller which is then slowly removed from the melt. If the temperature gradient of the melt is adjusted so that the melting/freezing temperature is just at the seed-melt interface, a continuous single crystal rod of silicon, called a **boule**, will grow as the puller is withdrawn.

[\[link\]](#) is a diagram of how the Czochralski process works. The entire apparatus must be enclosed in an argon atmosphere to prevent oxygen from getting into the silicon. The rod and the crucible are rotated in opposite directions to minimize the effects of convection in the melt. The pull-rate, the rotation rate and the temperature gradient must all be carefully optimized for a particular wafer diameter and growth direction. The $\langle 111 \rangle$ direction (along a diagonal of the cubic lattice structure) is usually chosen for wafers to be used for bipolar devices, while the $\langle 100 \rangle$ direction (along one of the sides of the cube) is favored for MOS applications. Currently, wafers are typically 6" or 8" in diameter, although 12" diameter wafers (300 mm) are looming on the horizon.



Czochralski crystal growth

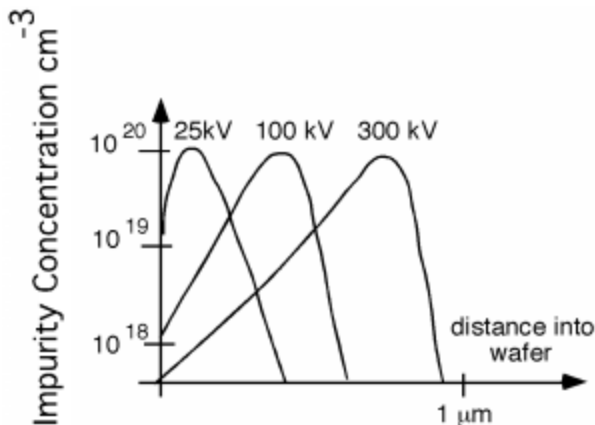
Once the boule is grown, it is ground down to a standard diameter (so the wafers can be used in automatic processing machines) and sliced into wafers, much like a salami. The wafers are etched and polished, and move on to the process line. A point to note however, is that due to "kerf" losses (the width of the saw blade) as well as polishing losses, more than half of the carefully grown, very pure, single crystal silicon is thrown away before the circuit fabrication process even begins!

Doping

Starting with a prepared, polished wafer then how do we get an integrated circuit? We will focus on the CMOS process, described in the last chapter. Let's assume we have wafer which was doped during growth so that it has a background concentration of acceptors in it (i.e. it is p-type). Referring back to [CMOS Logic](#), you can see that the first thing we need to build is a n-tank or moat. In order to do this, we need some way in which to introduce additional impurities into the semiconductor. There are several ways to do this, but current technology relies almost exclusively on a technique called **ion implantation**. A diagram of an ion-implanter is shown in the [figure in the previous section](#). An ion implanter uses a dopant source gas, ionizes it, and drives the ions into the wafer. The dopant gas is ionized and the resultant charged ions are accelerated through a magnetic field, where they are mass-analyzed. The vertical magnetic field causes the beam of ions to spread out, according to their mass. A thin aperture selects the ions of interest, and lets them pass, blocking all the others. This makes sure we are only implanting the ion we want, and in fact, even selects for the proper isotope! The ionized atoms are then accelerated through several tens to hundreds of kV, and then deflected by an electric field, much like in an oscilloscope CRT. In fact, most of the time the ion beam is "rastered" across the surface of the silicon wafer. The ions strike the silicon wafer and pass into its interior. A measurement of the current flow in the system and its integral, is a measure of how much dopant was deposited into the wafer. This is usually given in terms of the number of dopant $\frac{\text{atoms}}{\text{cm}^2}$ to which the wafer has been exposed.

After the atoms enter the silicon, they interact with the lattice, creating defects, and slowing down until finally they stop. Typical atomic distributions, as a function of implant voltage are show in [\[link\]](#) for implantation into amorphous silicon. When implantation is done on single crystal material, channeling, the improved mobility of an ion down the "hallway" of a given lattice direction, can skew the impurity distribution significantly. Just slight changes of less than a degree can make big differences in how the impurity atoms are finally distributed in the wafer. Usually, the operator of the implant machine purposely tilts the wafer a few

degrees off normal to the beam in order to arrive at more reproducible results.



Implant distribution with
acceleration energy

As you might expect, shooting 100 kV ions at a silicon wafer probably does quite a bit of damage to the crystal structure. Not only that, but just having, say boron, in your wafer does not mean you are going to have holes. For the boron to become "electrically active" - that is to act as an acceptor - it has to reside on a silicon lattice site. Even if the boron atom does, somehow, end up on an actual lattice site when it stops crashing around in the wafer, the many defects which have been created will act as deep traps. Thus, the hole which is formed will probably be caught at a trap site and will not be able to contribute to electrical conductivity in the wafer anyway. How can we fix this situation? If we carefully heat up the wafer, we can cause the atoms in the crystal to shake around, and if we do it right, they all get back where they belong. Not only that, but the newly added impurities end up on lattice sites as well! This step is called **annealing** and it does just what it is supposed to. Typical temperatures and times for such an anneal are 500 to 1000°C for 10 to 30 minutes.

Something else occurs during the anneal step however. We have just added, by our implantation step, impurities with a fairly tight distribution as shown in [\[link\]](#). There is an obvious gradient in impurity distribution, and if there

is a gradient, than things may start moving around by diffusion, especially at elevated temperatures.

Fick's First Law

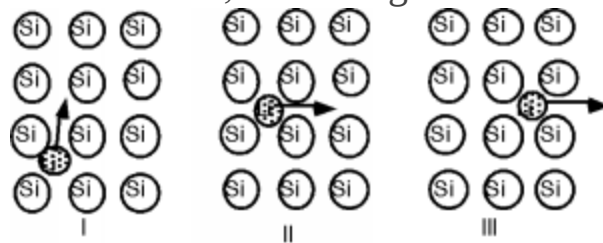
We talked about diffusion in the context of diodes, and described Fick's First Law of Diffusion for some particle concentration $N(x, t)$:

Fick's First Law of Diffusion

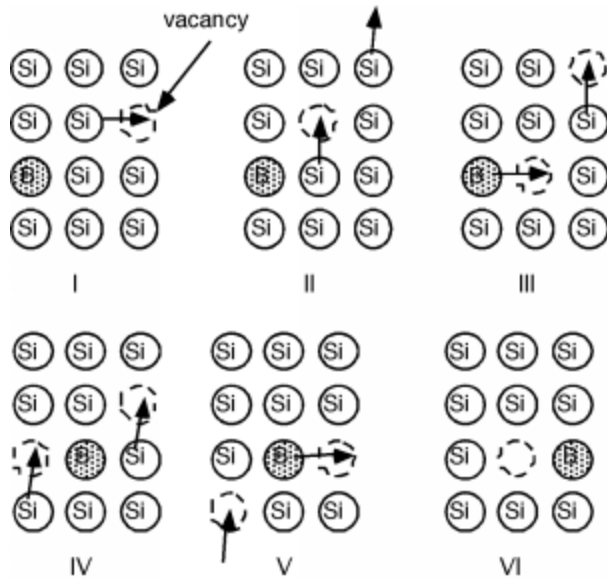
$$\text{Flux} = (-D) \frac{dN(x, t)}{dx}$$

D is the **diffusion coefficient** and has units of cm/sec.

In a semiconductor, impurities move about either **interstitially**, which means they travel around in-between the lattice sites ([\[link\]](#)). Or, they move by **substitutional diffusion**, which means they hop from lattice site to lattice site ([\[link\]](#)). Substitutional diffusion is only possible if the lattice has a number of **vacancies**, or empty lattice sites, scattered throughout the crystal, so that there are places into which the impurity can move. Moving interstitially requires energy to get over the potential barrier of the regions between the lattice sites. Energy is required to form the vacancies for substitutional diffusion. Thus, for either form of diffusion, the diffusion coefficient D , is a strong function of temperature.



Interstitial diffusion



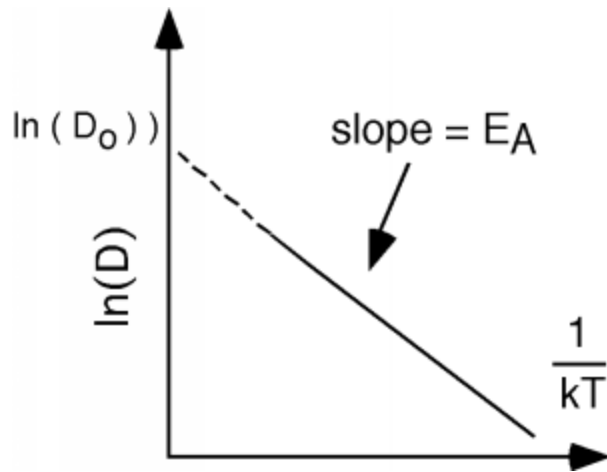
Substitutional diffusion

To a very good degree of accuracy, one can describe the temperature dependence of the diffusion coefficient with an **activation energy** E_A , such that:

Equation:

$$D(T) = D_o e^{-\frac{E_A}{kT}}$$

The activation energy E_A and coefficient D_o are obtained from a plot of the natural log of D vs. $\frac{1}{kT}$, called an **Arrhenius plot** ([\[link\]](#)). The slope gives E_A and the projection to infinite T ($\frac{1}{T} \rightarrow 0$) gives $\ln(D_o)$.



Arrhenius plot of diffusion
constant

The continuity equation holds for motion of impurities just like it does for anything else, so the divergence of the flux, $\text{div} (F)$ must equal the negative of the time rate of change of the concentration of the impurities, or, in one dimension:

Equation:

$$\frac{d}{d x}(\text{Flux}) = - \frac{d N(x, t)}{d t}$$

Fick's Second Law

Taking the derivative with respect to x of Fick's first law

Equation:

$$\frac{d}{dx}(\text{Flux}) = - \left(D \frac{\partial^2 N(x, t)}{\partial x^2} \right)$$

and then substituting the continuity equation into it, we have **Fick's second law of diffusion:**

Equation:

$$\frac{\partial N(x, t)}{\partial t} = D \frac{\partial^2 N(x, t)}{\partial x^2}$$

This is a standard diffusion equation, and one which shows up over and over again when one is dealing with such phenomena.

In order to get a solution to the diffusion equation, we must first assume some boundary conditions. We will deal with a semi-infinite wafer, and assume that

Equation:

$$\lim_{x \rightarrow \infty} N(x, t) = 0$$

This is a reasonable assumption, since at most our diffusion will only penetrate a micron or so into the wafer, and the whole wafer itself is several hundred microns thick.

We also have to decide something about initial conditions. We will make the assumption that we have at time $t = 0$ and $x = 0$ some surface concentration of impurities which we will call Q_0 ($\frac{\text{impurities}}{\text{cm}^2}$). This is the situation we would have if we introduce the impurities using a relatively shallow implant step. An alternative surface boundary condition would be

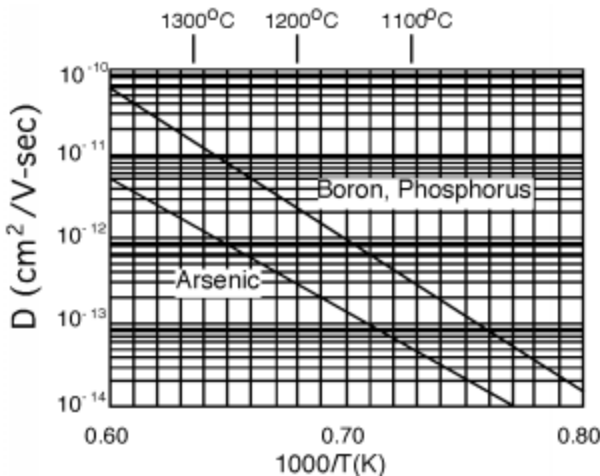
one where the concentration of impurities remains at some fixed value. This is what happens when there are impurities in the gas flow over the wafer during the time that they are in the diffusion oven. This is called an **infinite source diffusion**.

The first condition is called a **limited source diffusion** and that is what we shall consider further here. It is not too hard to show that with this initial condition, the solution to the diffusion equation is:

Equation:

$$N(x, t) = \frac{Q_0}{\sqrt{\pi Dt}} e^{-\frac{x^2}{4Dt}}$$

Note that $N(x, t)$ is a function of distance into the wafer, and time t . The time is, of course, the time of the diffusion process. D , the diffusion constant, depends on the temperature at which the diffusion takes place. [\[link\]](#) is a plot of D for three of the most commonly used dopants in silicon. Phosphorus and boron are the most common acceptor and donor respectively. Arsenic is sometimes used because it is significantly bigger in diameter than either P or B and thus, moves around less after an implant.



Diffusion constant as a function
of 1000/T

Suppose we do a relatively shallow implant of boron into our p-type wafer, and deposit a Q_0 of 5×10^{13} phosphorus $\frac{\text{atoms}}{\text{cm}^2}$. We then perform an anneal diffusion at 1100°C for 60 minutes. At 1100°C , D for phosphorus seems to be about $2 \times 10^{-13} \frac{\text{cm}^2}{\text{sec}}$. We will make a plot of $N(x)$ for various times. If you do this at home, be sure to put time in seconds, not minutes, hours, or fortnights. Looking at [\[link\]](#), is pretty easy to see how the impurities move into the semiconductor, and how the concentration at the surface, $N(0, t)$, decreases as more and more of the impurities moves deeper into the wafer.

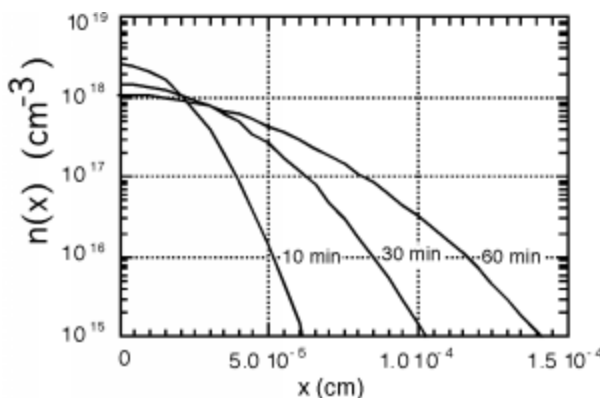
Exercise:

Problem:

If the substrate had been doped at $10^{16} \frac{\text{acceptors}}{\text{cm}^3}$ where would be the location of the p-n junction between the implanted phosphorus layer, and the background boron?

Solution:

About $1.2 \mu\text{m}$ after 1 hour of diffusion time. You know this because for $x < 1.2 \mu\text{m}$ the phosphorus concentration is greater than that of boron, and so the material is n-type. For $x > 1.2 \mu\text{m}$, the boron concentration exceeds that of the phosphorous, and so the material is now p-type.)



Diffusion distribution at
different times

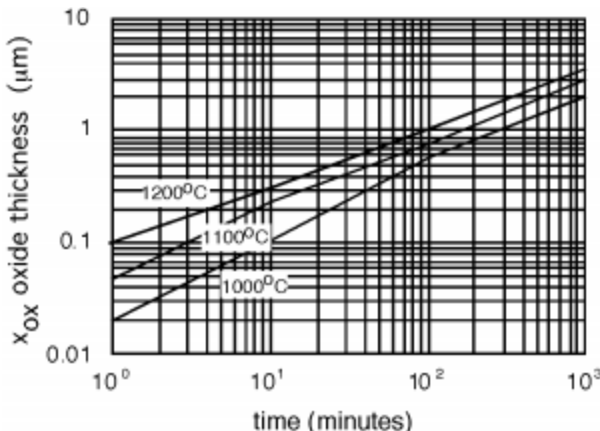
Photolithography

Actually, implants (especially for moats) are usually done at a sufficiently high energy so that the dopant (phosphorus) is already pretty far into the substrate (often several microns or so), even before the diffusion starts. The anneal/diffusion moves the impurities into the wafer a bit more, and as we shall see also makes the n-region grow larger.

"The n-region"! We have not said a thing about how we make our moat in only certain areas of the wafer. From the description we have so far, it seems we have simply built an n-type layer over the whole surface of the wafer. This would be bad! We need to come up with some kind of "window" to only permit the implanting impurities to enter the silicon wafer where we want them and not elsewhere. We will do this by constructing an implantation "barrier".

To do this, the first thing we do is grow a layer of silicon dioxide over the entire surface of the wafer. We talked about oxide growth when we were discussing MOSFETs but let's go into a little more detail. You can grow oxide in either a dry oxygen atmosphere, or in an atmosphere which contains water vapor, or steam. In [\[link\]](#), we show oxide thickness as a function of time for growth with steam. Dry O_2 does not behave too much differently, the rate is just somewhat slower.

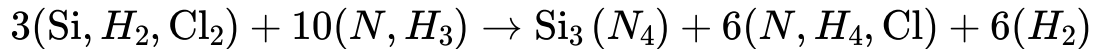
Oxide Thickness as a Function of Time



On top of the oxide, we are now going to deposit yet another material. This is silicon nitride, $Si_3 N_4$ or just plain "**nitride**" as it is usually called. Silicon nitride is deposited through a method called chemical vapor deposition or

"CVD". The usual technique is to react dichlorosilane and ammonia in a hot walled low pressure chemical vapor deposition system (LPCVD). The reaction is:

Equation:



Silicon nitride is a good barrier for impurities, oxygen and other things which do not want to get into the wafer. Take a look at [\[link\]](#) and see what we have so far. A word about scale and dimensions. The silicon wafer is about 250µm thick (about 0.01") since it has to be strong enough not to break as it is being handled. The two deposited layers are each about 1µm thick, so they should actually be drawn as lines thinner than the other lines in the figure. This would obviously make the whole idea of a sketch ridiculous, so we will leave things distorted as they are, keeping in mind that the deposited and diffused layers are actually **much** thinner than the rest of wafer, which really does not do anything except support the active circuits up on top. (There we go again, wasting silicon. Good thing it's cheap and plentiful!)

Initial Wafer Configuration



Now what we want to do is remove **part** of the nitride, so we can make our n-well, but not put in phosphorous where do not want it. We do this with a processes called **photolithography** and **etching** respectively. First thing we do is coat the wafer with yet another layer of material. This is a liquid called **photoresist** and it is applied through a process called **spin-coating**. The wafer is put on a vacuum chuck, and a layer of liquid photoresist is sprayed uncap of the wafer. The chuck is then spun rapidly, getting to several thousand RPM in a small fraction of a second. Centrifugal force causes the resist to spread out uniformly across the wafer surface (most of it

in fact flies off!). The solvent for the photoresist is quite volatile and so the layer of photoresist dries while the wafer is still spinning, resulting in a thin, uniform coating across the wafer [\[link\]](#).

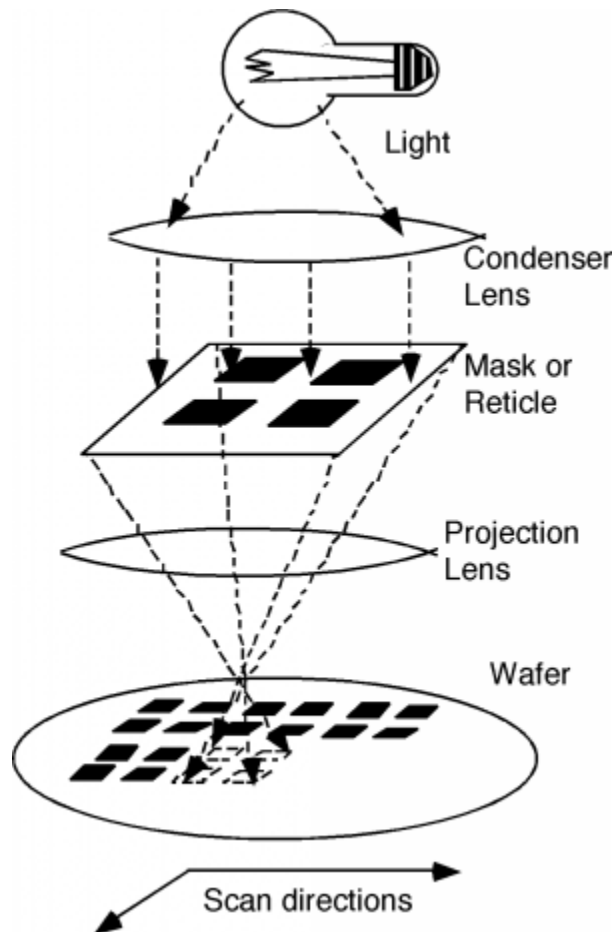
Photoresist is Spun On



The name "photoresist" gives some clue as to what this stuff is. Basically, photoresist is a polymer mixed with some kind of light sensitizing compound. In **positive** photoresist, wherever light strikes it, the polymer is weakened, and it can be more easily removed with a solvent during the **development** process. Conversely, **negative** photoresist is strengthened when it is illuminated with light, and is more resistant to the solvent than is the unilluminated photoresist. Positive resist is so-called because the image of the developed photoresist on the wafer looks just like the mask that was used to create it. Negative photoresist makes an image which is the opposite of what the mask looks like.

We have to come up with some way of selectively illuminating certain portions of the photoresist. Anyone who has ever seen a projector know how we can do this. But, since we want to make **small** things, not big ones, we will change around our projector so that it makes a smaller image, instead of a bigger one. The instrument that projects the light onto the photoresist on the wafer is called a **projection printer** or a **stepper** [\[link\]](#).

A Stepper Configuration

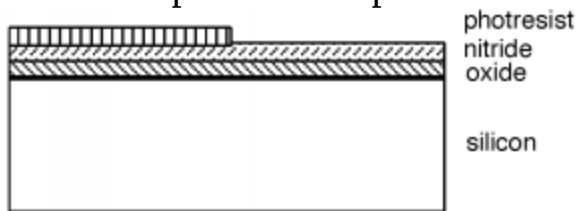


As shown in [\[link\]](#), the stepper consists of several parts. There is a light source (usually a mercury vapor lamp, although ultra-violet excimer lasers are also starting to come into use), a condenser lens to image the light source on the **mask** or **reticle**. The mask contains an image of the **pattern** we are trying to place on the wafer. The projection lens then makes a reduced (usually 5x) image of the mask on the wafer. Because it would be far too costly, if not just plain impossible, to project onto the whole wafer all at once, only a small selected area is printed at one time. Then the wafer is **scanned** or **stepped** into a new position, and the image is printed again. If previous patterns have already been formed on the wafer, TV cameras, with artificial intelligence algorithms are used to **align** the current image with the previously formed features. The stepper moves the whole surface of the wafer under the lens, until the wafer is completely covered with the desired pattern. A stepper is not cheap. Usually, TI or Intel will fork over several million dollars for each one! It is one of the most important pieces

of equipment in the whole IC fab however, since it determines what the minimum feature size on the circuit will be.

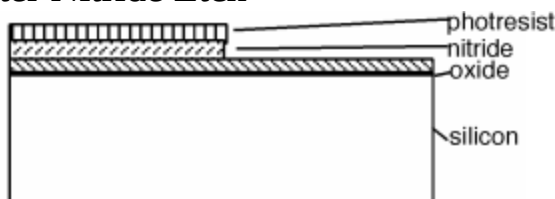
After exposure, the photoresist is placed in a suitable solvent, and "developed". Suppose for our example the structure shown in [\[link\]](#) is what results from the photolithographic step.

After PR Expose/Develop



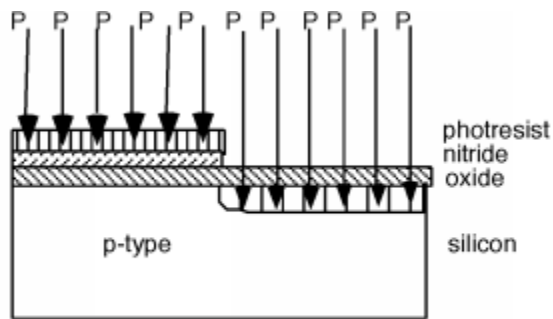
The pattern that was used in the photolithographic (PL) step exposed half of our area to light, and so the photoresist (PR) in that region was removed upon development. The wafer is now immersed in a hydrofluoric acid (HF) solution. HF acid etches silicon nitride quite rapidly, but does not etch silicon dioxide nearly as fast, so after the etch we have what we see in [\[link\]](#).

After Nitride Etch



We **now** take our wafer, put it in the ion implanter and subject it to a "blast" of phosphorus ions [\[link\]](#).

Implanting Phosphorus

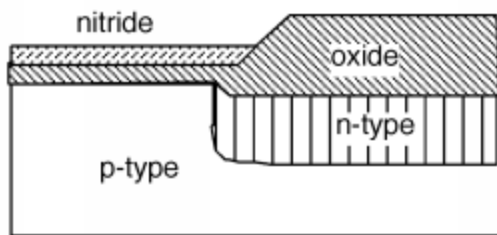


The ions go right through the oxide layer on the RHS, but stick in the resist/nitride layer on the LHS of our structure.

Integrated Circuit Well and Gate Creation

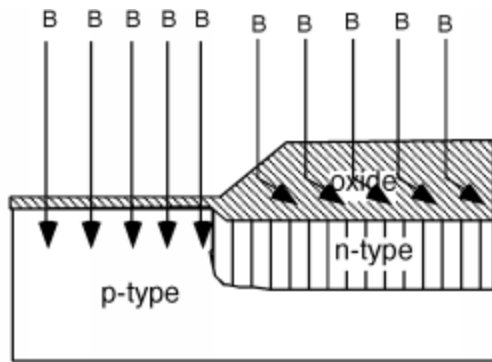
We then remove the remaining resist, and perform an activation/anneal/diffusion step, also sometimes called the "drive-in". The purpose of this step is two fold. We want to make the n-tank deep enough so that we can use it for our p-channel MOS, and we want to build up an implant barrier so that we can implant into the p-substrate region only. We introduce oxygen into the reactor during the activation, so that we grow a thicker oxide over the region where we implanted the phosphorus. The nitride layer over the p-substrate on the LHS protects that area from any oxide growth. We then end up with the structure shown in [\[link\]](#).

After the Anneal/Drive-In



Now we strip the remaining nitride. Since the only way we can convert from p to n is to add a donor concentration which is **greater** than the background acceptor concentration, we had to keep the doping in the substrate fairly light in order to be able to make the n-tank. The lightly doped p-substrate would have too low a threshold voltage for good n-MOS transistor operation, so we will do a V_T adjust implant through the thin oxide on the LHS, with the thick oxide on the RHS blocking the boron from getting into the n-tank. This is shown in [\[link\]](#), where boron is implanted into the p-type substrate on the LHS, but is blocked by the thick oxide in the region over the n-well.

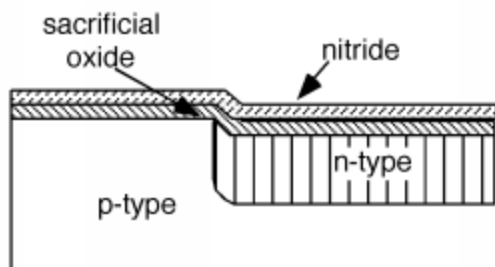
Adjust Implant



V_T adjust implant

Next, we strip off all the oxide, grow a new thin layer of oxide, and then a layer of nitride [\[link\]](#). The oxide layer is grown only because it is bad to grow Si_3N_4 directly on top of silicon, as the different coefficients of thermal expansion between the two materials causes damage to the silicon crystal structure. Also, it turns out to be nearly impossible to remove nitride if it is deposited directly on to silicon.

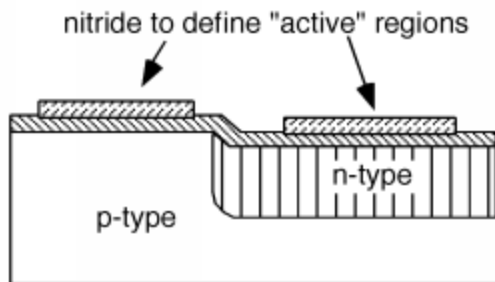
Strip Oxide, New Nitride



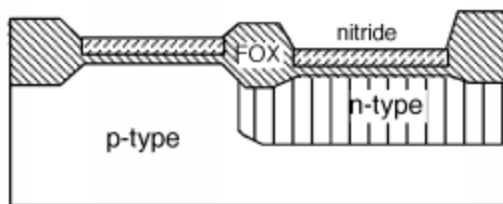
The nitride is patterned (covered with photoresist, exposed, developed, etched, and removal of photoresist) to make two areas which are called "active" [\[link\]](#). (This is where we will build our transistors.) The wafer is then subjected to a high-pressure oxidation step which grows a thick oxide wherever the nitride was removed. The nitride is a good barrier for oxygen, so essentially no oxide grows underneath it. The thick oxide is used to isolate individual transistors, and also to make for an insulating layer over

which conducting patterns can be run. The thick oxide is called **field oxide** (or FOX for short) [\[link\]](#).

Nitride After Etching

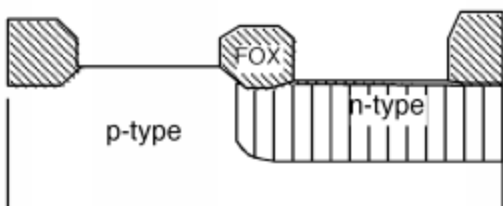


Growing Field Oxide



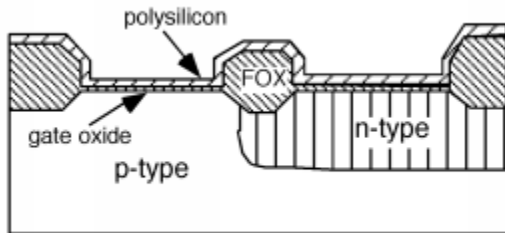
Then, the nitride, and some of the oxide are etched off. The oxide is etched enough so that all of the oxide under the nitride regions is removed, which will take a little off the field oxide as well. This is because we now want to grow the gate oxide, which must be very clean and pure [\[link\]](#). The oxide under the nitride is sometimes called **sacrificial** oxide, because it is sacrificed in the name of ultra performance.

Ready to Grow Gate Oxide



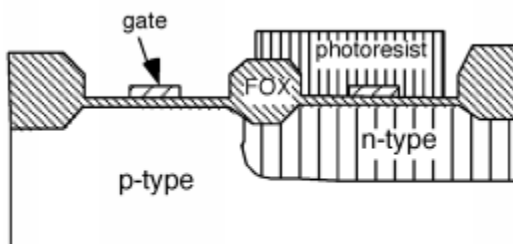
Then the gate oxide is grown, and immediately thereafter, the whole wafer is covered with polysilicon [\[link\]](#).

Poly Deposition Over Gate Oxide



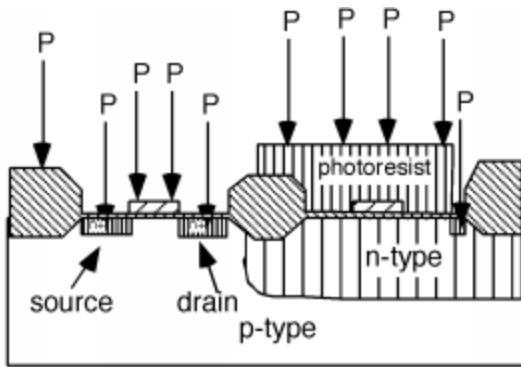
The polysilicon is then patterned to form the two regions which will be our gates. The wafer is covered once again with photoresist. The resist is removed over the region that will be the n-channel device, but is left covering the p-channel device. A little area near the edge of the n-tank is also uncovered [\[link\]](#). This will allow us to add some additional phosphorus into the n-well, so that we can make a contact there, so that the n-well can be connected to V_{dd} .

Preparing for NMOS Channel/Drain Implant



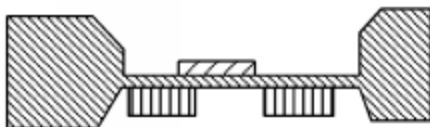
Back into the implanter we go, this time exposing the wafer to phosphorus. The poly gate, the FOX and the photoresist all block phosphorus from getting into the wafer, so we make two n-type regions in the p-type substrate, and we have made our n-channel MOS source/drain regions. We also add phosphorus to the V_{dd} contact region in the n-well so as to make sure we get good contact performance there [\[link\]](#).

Phosphorus S/D Implant

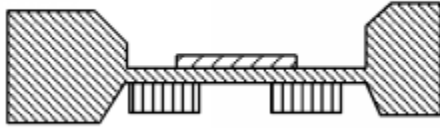


Note that the formation of the source and drain were performed with a **self-aligning technology**. This means that we used the gate structure itself to define where the two inside edges of the source and drain would be for the MOSFET. If we had made the source/drain regions **before** we defined the gate, and then tried to line the gate up right over the space between them, we might have gotten something that looks like what is shown in [\[link\]](#). What's going to be the problem with this transistor? Obviously, if the gate does not extend all the way to both the source **and** the drain, then the channel will not either, and the transistor will never turn on! We could try making the gate wider, to ensure that it will overlap both active areas, even if it is slightly misaligned, but then you get a lot of extraneous fringing capacitance which will significantly slow down the speed of operation of the transistor [\[link\]](#). This is bad! The development of the self-aligned gate technique was one of the big breakthroughs which has propelled us into the VLSI and ULSI era.

Misaligned Gate

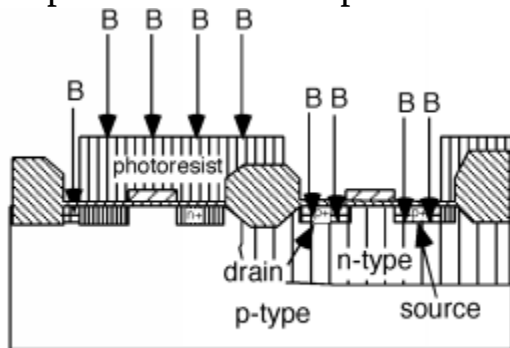


Wide Gate



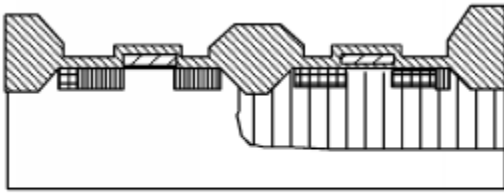
We pull the wafer out of the implanter, and strip off the photoresist. This is sometimes difficult, because the act of ion implantation can "bake" the photoresist into a very tough film. Sometimes an rf discharge in an O_2 atmosphere is used to "ash" the photoresist, and literally burn it off the wafer! We now apply some more PR, and this time pattern to have the moat area, and a substrate contact exposed, for a boron p+ implant. This is shown in [\[link\]](#).

Boron p-Channel S/D Implant

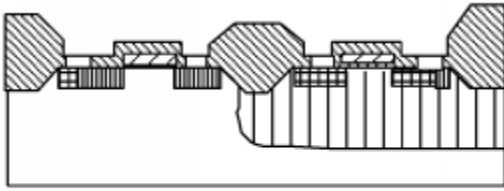


We are almost done. The next thing we do is remove all the photoresist, and grow one more layer of oxide, which covers everything, as shown in [\[link\]](#). We put photoresist over the whole wafer again, and pattern for contact holes to go through the oxide. We will put contacts for the two drains, and for each of the sources, make sure that the holes are big enough to also allow us to connect the source contact to either the p-substrate or the n-moat as is appropriate [\[link\]](#).

Final Oxide Growth



Contact Holes Etched

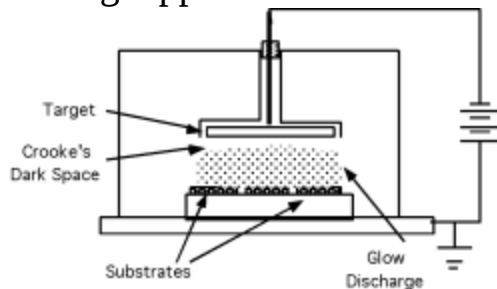


Applying Metal/Sputtering

We now put the wafer in a **sputter deposition system**. In the sputter system, we coat the entire surface of the wafer with a conductor. An aluminum-silicon alloy is usually used, although other metals are employed as well.

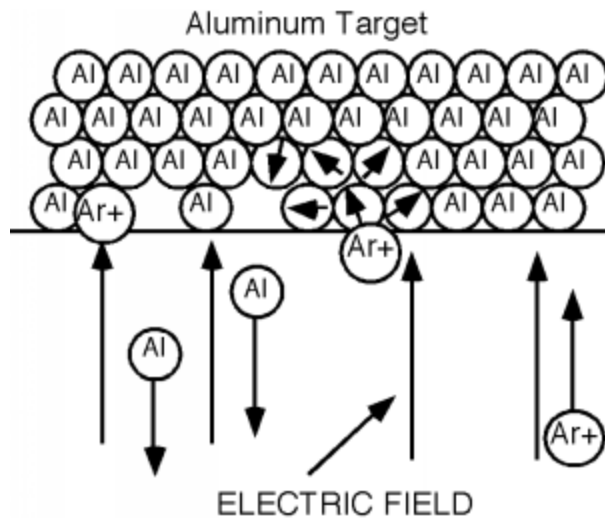
A sputtering system is shown schematically in [\[link\]](#). A sputtering system is a vacuum chamber, which after it is pumped out, is re-filled with a low-pressure argon gas. A high voltage ionizes the gas, and creates what is known as the **Crookes dark space** near the cathode, which in our case, consists of a metal target made out of the metal we want to deposit. Almost all of the potential of the high-voltage supply appears across the dark space. (The glow discharge consists of argon ions and electrons which have been stripped off of them. Since there are about equal number of ions and electrons, the net charge density is about zero, and hence by Gauss' law, so is the field.)

Sputtering Apparatus

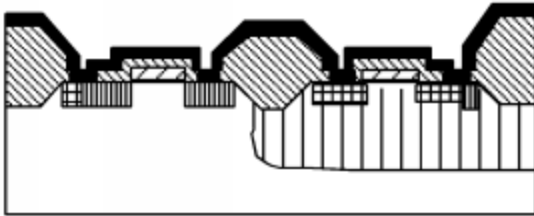


The electric field accelerates the argon atoms which slam into the aluminum target. There is an exchange of momentum, and an aluminum atom is ejected from the target ([\[link\]](#)) and heads to the silicon wafer, where it sticks, and builds up a metal film [\[link\]](#).

Sputtering Mechanism

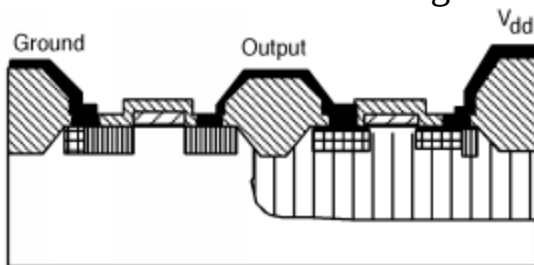


Wafer Coated with Metal



If you look at [\[link\]](#), you will note that we have seemingly done something pretty stupid. We have wired all of the elements of our CMOS inverter together! Ah, but all is not lost. We can do one more photolithographic step, and pattern and etch the aluminum, so we only have it where we need it. This is shown in [\[link\]](#).

After Interconnect Patterning



Integrated Circuit Manufacturing: Bird's Eye View

It will no doubt be helpful if we also take a plane or "bird's eye" view of what this circuit looks like as well. There are, in fact, some interesting things we can gain by looking at some of them.

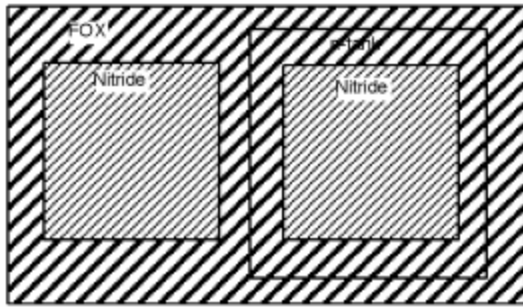
We have been looking at the development of the circuit from a cross-sectional point of view, watching the formation of the various levels which make up the finished CMOS inverter. This is, in fact, not the way a circuit designer looks at things. A circuit designer sees things from above, and only worries about the placement of transistors, and how they will be connected together. In fact, the only factor in the actual design of the layout engineer has any choice on is the transistor width, W . All other parameters are decided upon beforehand by the process engineer. So what does the layout engineer see? We start with the n- implant to make the n-tank, as shown in [\[link\]](#). (You should go back and follow along with the cross-sectional views of the process, as we review looking at things from the top.)

Implanted n-Tank

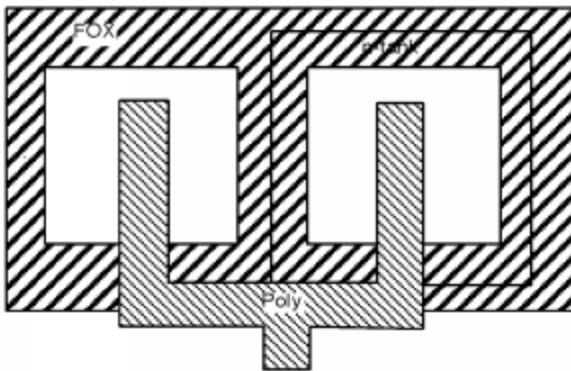


A mask opposite to that of the n-tank allows us to an n-channel V_T adjust. We next deposit and pattern the nitride for the active regions, and grow the field oxide (FOX) [\[link\]](#).

Growing FOX

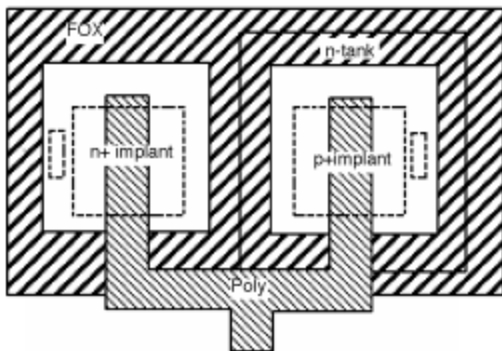


We remove the nitride, and deposit and pattern the poly., as seen in [\[link\]](#)
 Gate Poly Pattern



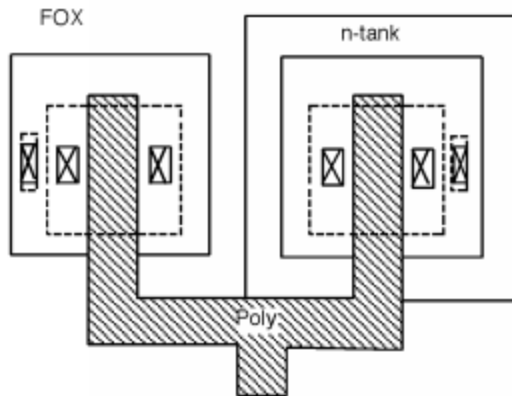
[\[link\]](#) shows what the two masks look like for the n⁺ and p⁺ source/drain implants:

S/D Implants



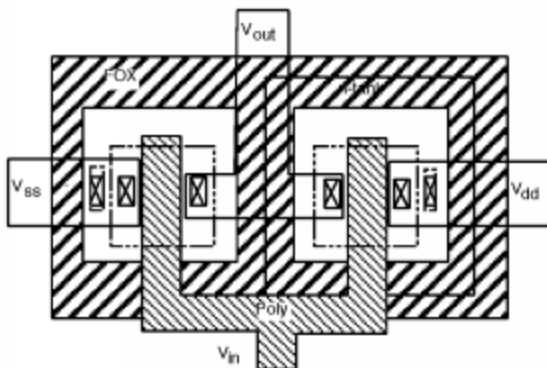
Note that the gate poly extends beyond where the implant is being performed (inside the dotted line). This is a **design rule** which is the way the circuit designer takes into account the fact that the manufacturing

process must have some tolerance built in, because things will not always be lined up just perfectly. Now we make some contact holes, seen in [\[link\]](#):
Etching Contact Holes



And finally, we sputter and pattern the metallization, which is depicted in [\[link\]](#). You should go back to [MOSFETs](#), and convince yourself that the circuit shown in [\[link\]](#) is indeed what has been constructed in [\[link\]](#). See if you can identify all of the correct parts. Note that there is a connection between V_{ss} (ground) and the p-substrate **very** close to the n-channel source. There is also a contact between the n-moat and V_{dd} which is **very** close to the p-channel source. What advantage would this have? Hint: review the discussion of [latch-up](#).

Metallization Patterning



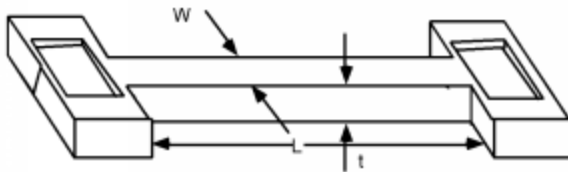
Diffused Resistor

Sometimes, in a circuit design, we will need a resistor. This is usually made either with poly or with a [diffusion](#). If we took our n-tank or similar n-type diffusion, we could make a long narrow strip of it, and use it as a resistor. As long as we keep the substrate at ground, and any voltages on the resistor greater than ground, the n-p junction will be reverse biased and the resistor will be isolated from the substrate. Now we all know

Equation:

$$\begin{aligned} R &= \frac{\rho L}{A} \\ &= \frac{L}{nq\mu tW} \end{aligned}$$

A Diffused Resistor



The only trouble is, what is n for a diffused resistor? A quick look at the [chart](#) showing carrier concentration as a function of depth after a diffusion shows that when we do a diffusion, n is not a constant, but varies as we go down into the wafer. We will have to do some kind of integral, assuming lots of parallel, thin resistors, each with a different carrier concentration! This is not very satisfactory.

In fact, it is so unsatisfactory that IC engineers have come up with a better description resistance than one involving n and μ . Note that we could write [\[link\]](#) as

Equation:

$$R = \frac{1}{nq\mu t} \frac{L}{W}$$

We define the first fraction (which contains the carrier concentration, thickness etc.) as the **sheet resistance** R_s of the diffusion. While this can be more-or-less predicted, it is usually also a post-fabrication measured value.

Equation:

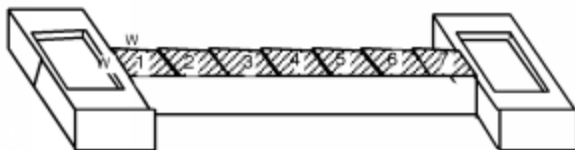
$$R_s \equiv \frac{1}{nq\mu t}$$

R_s has units of "ohms/square", and you are probably tempted to ask "per square what?". Well it can be any square at all, cm, μm , km, since all we really need to know is R_s and the length to width ratio of the resistor structure to find the resistance of a resistor. We do not need to know what units are used to measure the length and the width, so long as they are the same for both. For instance if the resistor in [\[link\]](#) has a sheet resistivity of 50 Ω/square , then by blocking the resistor off into squares $W \times W$ in dimension, we see that the resistor is 7 squares long ([\[link\]](#)) and so its resistance is given as:

Equation:

$$\begin{aligned} R &= 50 \left(\frac{\Omega}{\text{square}} \right) 7(\text{squares}) \\ &= 350(\Omega) \end{aligned}$$

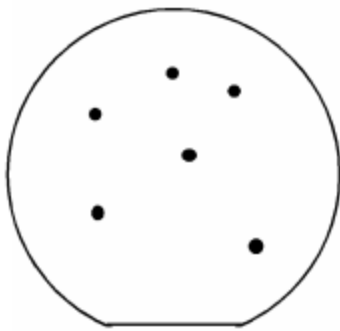
Counting the Squares



Yield

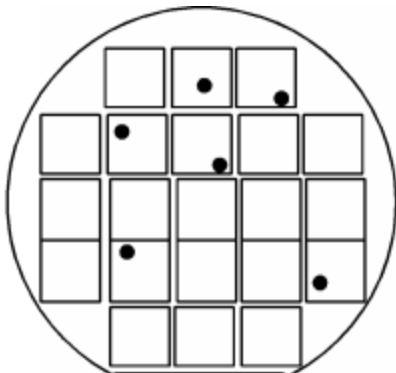
Perhaps a word about feature size, chip size and yield would be in order. We saw earlier that circuits are repeated many times across a wafer's surface during the photolithographic stage. Although great care is exercised in trying to prevent defects from becoming part of a wafer surface (clean rooms, "bunny" suits, ultra-pure chemicals etc.) each wafer that goes through a fab will end up with **some** "killer" defects distributed across the wafer surface [\[link\]](#).

A Wafer with Defects

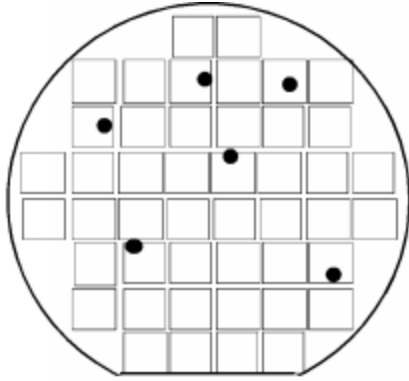


Imagine that we try to manufacture some chips of a certain size. A glance at [\[link\]](#) shows that we would have 15 of 21 good chips, for a yield of about 71%. Suppose we could, through improved technology, perform a 30% "shrink" on the circuit - i.e. make its dimensions 30% smaller. Now, as [\[link\]](#) shows, we get 40 good chips/wafer instead of 15 (and they cost no more to produce) and our yield has gone to 40 out of 46 or 87%. We will be rich! Or at least we won't go out of business!

Six Killed Circuits



Lots More Good Ones



Yield, reliability and manufacturability are all critical issues in the semiconductor industry. The business is highly competitive, and the technology keeps moving rapidly. It is an exciting and challenging field, one which demands the very best, but which rewards someone who is willing to never stop thinking and to bring forth the very best creative solutions to hard problems.

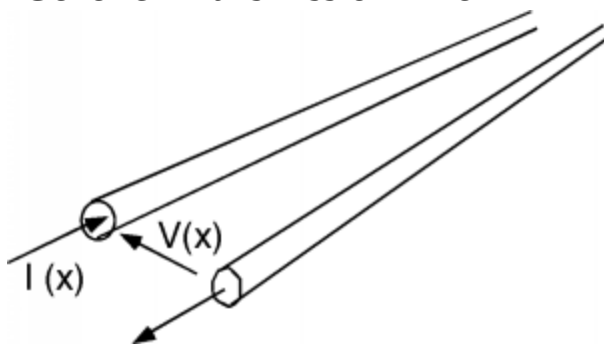
Distributed Parameters

Having learned something about how we generate signals with bipolar and field effect transistors, we now turn our attention to the problem of getting those signals from one place to the next. Ever since Samuel Morse (and the founder of **my alma mater**, Ezra Cornell) demonstrated the first working telegraph, engineers and scientists have been working on the problem of describing and predicting how electrical signals behave as they travel down specific structures called **transmission lines**.

Any electrical structure which carries a signal from one point to another can be considered a transmission line. Be it a long-haul coaxial cable used in the Internet, a twisted pair in a building as part of a local-area network, a cable connecting a PC to a printer, a bus layout on a motherboard, or a metallization layer on an integrated circuit, the fundamental behavior of all of these structures are described by the same basic equations. As computer switching speeds run into the 100s of MHz, into the GHz range, considerations of transmission line behavior are ever more critical, and become a more dominant force in the performance limitations of any system.

For our initial purposes, we will introduce a "generic" transmission line [\[link\]](#), which will incorporate most (but not all) features of real transmission lines. We will then make some rather broad simplifications, which, while rendering our results less applicable to real-life situations, nevertheless **greatly** simplify the solutions, and lead us to insights that we can indeed apply to a broad range of situations.

"Generic" Transmission Line



The generic line consists of two conductors. We will suppose a potential difference $V(x)$ exists between the two conductors, and that a current $I(x)$ flows down one conductor, and returns via the other. For the time being, we will let the transmission line be "semi-infinite", which means we have access to the line at some point x , but the line then extends out in the x direction to infinity. (Such lines are a bit difficult to handle in the lab!)

In order to be able to describe how $V(x)$ and $I(x)$ behave on this line, we have to make some kind of **model** of the electrical characteristics of the line itself. We can not just make up any model we want however; we have to base the model on physical realities.

Let's start out by just considering one of the conductors and the physical effects of current flowing through that conductor. We know from freshman physics that a current flowing in a wire gives rise to a magnetic field, H ([\[link\]](#)). Multiply H by μ and you get B , the magnetic flux density, and then integrate B over a plane parallel to the wires and you get Φ , the magnetic flux "linking" the circuit. This is shown in [\[link\]](#) for at least part of the surface. The definition of L , the inductance of a circuit element, is just

Equation:

$$L \equiv \frac{\Phi}{I}$$

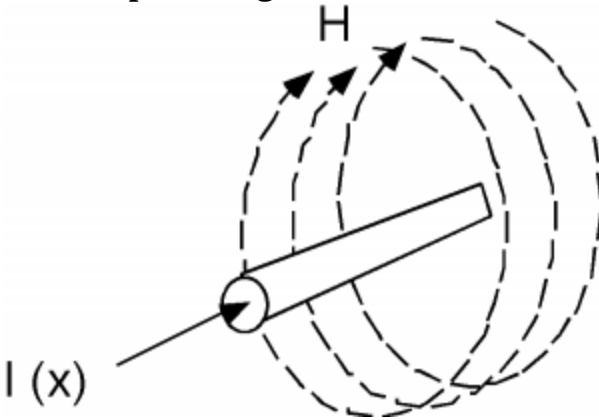
where Φ is the flux linking the circuit element, and I is the current flowing through it. Our only problem in finding Φ is that the longer a section of wire we take, the more Φ we have for the same I . Thus, we will introduce the concept of a distributed parameter.

distributed parameter

A distributed parameter is a parameter which is spread throughout a structure and is not confined to a lumped element such as a coil of wire.

Example:

For instance, we will hereby define \mathbf{L} as the **distributed inductance** for the transmission line. It has units of Henrys/meter. If we have a length of transmission line x_0 meters long, and if that line has a distributed inductance of \mathbf{L} H/m, then the inductance L of that length of line is just $L = \mathbf{L}x_0$.

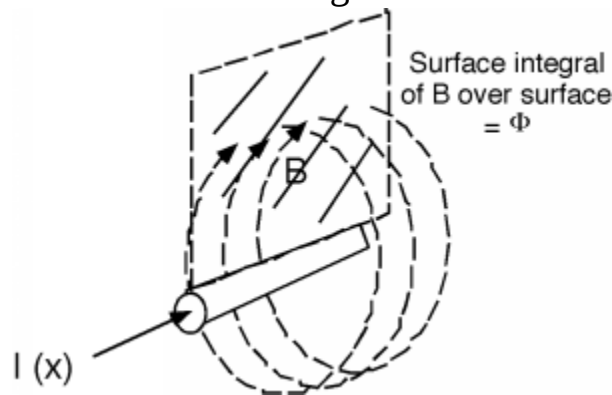
Build Up of Magnetic Field

Likewise, if we have two conductors separated by some distance, and if there is a potential difference V between the conductors, then there must be some charge $\pm(Q)$ on the two conductors which gives rise to that potential difference. We can imagine a linear charge distribution on the transmission line, ρ (C/m), where we have ρ Coulombs/m on one conductor, and $-\rho$ Coulombs/m on the other conductor. For a line of length x_0 , we would have $Q = \pm(\rho x_0)$ on each section of wire. Whenever you have two charged conductors with a voltage difference between them, you can describe the ratio of the charge to the voltage as a capacitance. The two conductors would have a capacitance

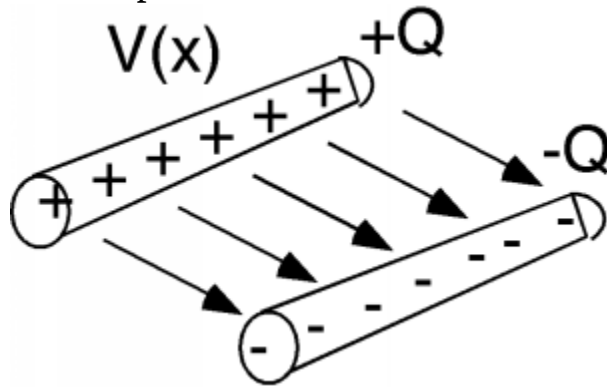
Equation:

$$\begin{aligned} C &= \frac{Q}{V} \\ &= \frac{\rho x_0}{V} \end{aligned}$$

and a distributed capacitance C (F/m) which is just $\frac{\rho}{V}$. A length of line x_0 long would have a capacitance $C = Cx_0$ Farads associated with it [\[link\]](#).
Find the Flux Linkage



Line Capacitance

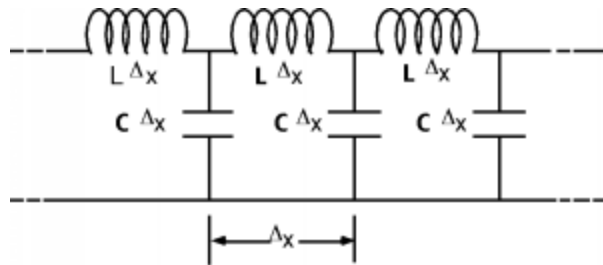


Thus, we see that the transmission line has both a distributed inductance L and a distributed capacitance C which are tied up with each other. There is really no way in which we can separate one from the other. In other words, we can not have only the capacitance, or only the inductance, there will always be some of each associated with each section of line now matter how small or how big we make it.

We are now ready to build our model. What we want to do is to come up with some arrangement of inductors and capacitors which will represent electrically, the properties of the distributed capacitance and inductance we discussed above. As a length of line gets longer, its capacitance increases, so we had better put the distributed capacitances in parallel with one another, since that is the way capacitors add up. Also, as the line gets longer, its total inductance increases, so we had better put the distributed inductances in series with one another, for that is the way inductances add

up. [\[link\]](#) is a representation of the distributed inductance and capacitance of the generic transmission line.

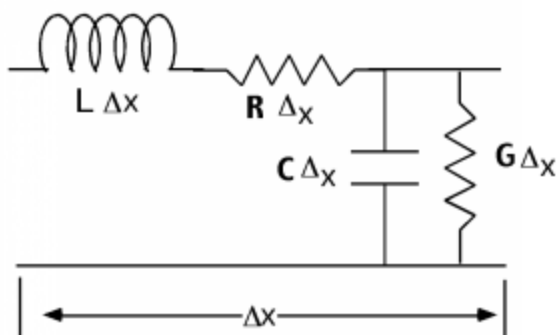
Distributed Parameter Model



We break the line up into sections $\Delta(x)$ long, each one with an inductance $L\Delta(x)$ and a capacitance $C\Delta(x)$. If we halve $\Delta(x)$, we would halve the inductance and capacitance of each section, but we'd have twice as many of them per unit length. Duh! The point is no matter how fine we make $C\Delta(x)$, we still have Ls and Cs arranged like we see in [\[link\]](#), with the two kinds of components intermixed.

We **could** make a more realistic model and realize that all real wires have series resistance associated with them and that whatever we use to keep the two conductors separated will have some leakage conductance associated it. To account for this we would introduce a series resistance R (ohms/unit length) and a series conductance G (ohms/unit length). One section of our line model then looks like [\[link\]](#).

Complete Distributed Model



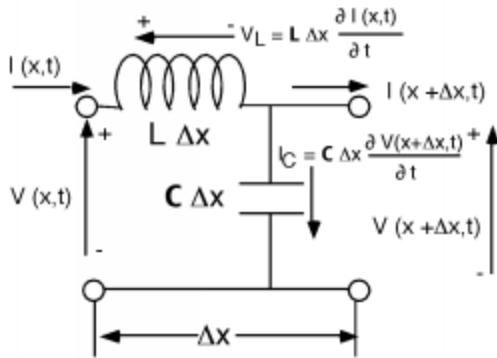
Although this **is** a more realistic model, it leads to much more complicated math. We will start out anyway, ignoring the series resistance R and the shunt conductance G . This "approximation" turns out to be pretty good as long as either the line is not too long, or the frequencies of the signals we

are sending down the line do not get too high. Without the series resistance or parallel conductance we have what is called an ideal **lossless transmission line**.

Telegrapher's Equations

Let's look at just one little section of the line, and define some voltages and currents [\[link\]](#).

Applying Kirchhoff's Laws



For the section of line $\Delta(x)$ long, the voltage at its input is just $V(x, t)$ and the voltage at the output is $V(x + \Delta(x), t)$. Likewise, we have a current $I(x, t)$ entering the section, and another current $I(x + \Delta(x), t)$ leaving the section of line. Note that both the voltage and the current are functions of **time** as well as position.

The voltage drop across the inductor is just:

Equation:

$$V_L = L\Delta(x) \frac{\partial I(x, t)}{\partial t}$$

Likewise, the current flowing down through the capacitor is

Equation:

$$I_C = C\Delta(x) \frac{\partial V(x + \Delta(x), t)}{\partial t}$$

Now we do a [KVL](#) around the outside of the section of line and we get

Equation:

$$V(x, t) - V_L - V(x + \Delta(x), t) = 0$$

Substituting [\[link\]](#) for V_L and taking it over to the RHS we have

Equation:

$$V(x, t) - V(x + \Delta(x), t) = \mathbf{L}\Delta(x) \frac{\partial I(x, t)}{\partial t}$$

Let's multiply by -1, and bring the $\Delta(x)$ over to the left hand side.

Equation:

$$\frac{V(x + \Delta(x), t) - V(x, t)}{\Delta(x)} = - \left(\mathbf{L} \frac{\partial I(x, t)}{\partial t} \right)$$

We take the limit as $\Delta(x) \rightarrow 0$ and the LHS becomes a derivative:

Equation:

$$\frac{\partial V(x, t)}{\partial x} = - \left(\mathbf{L} \frac{\partial I(x, t)}{\partial t} \right)$$

Now we can do a [KCL](#) at the node where the inductor and capacitor come together.

Equation:

$$I(x, t) - \mathbf{C}\Delta(x) \frac{\partial V(x + \Delta(x), t)}{\partial t} - I(x + \Delta(x), t) = 0$$

And upon rearrangement:

Equation:

$$\frac{I(x + \Delta(x), t) - I(x, t)}{\Delta(x)} = - \left(\mathbf{C} \frac{\partial V(x + \Delta(x), t)}{\partial t} \right)$$

Now when we let $\Delta(x) \rightarrow 0$, the left hand side again becomes a derivative, and on the right hand side, $V(x + \Delta(x), t) \rightarrow V(x, t)$, so we have:

Equation:

$$\frac{\partial I(x, t)}{\partial x} = - \left(C \frac{\partial V(x, t)}{\partial t} \right)$$

[\[link\]](#) and [\[link\]](#) are so important we will write them out again together:

Equation:

$$\frac{\partial V(x, t)}{\partial x} = - \left(L \frac{\partial I(x, t)}{\partial t} \right)$$

Equation:

$$\frac{\partial I(x, t)}{\partial x} = - \left(C \frac{\partial V(x, t)}{\partial t} \right)$$

These are called the **telegrapher's equations** and they are all we really need to derive how electrical signals behave as they move along on transmission lines. Note what they say. The first one says that at some point x along the line, the incremental voltage drop that we experience as we move down the line is just the distributed inductance L times the time derivative of the current flowing in the line at that point. The second equation simply tells us that the loss of current as we go down the line is proportional to the distributed capacitance C times the time rate of change of the voltage on the line. As you should be easily aware, what we have here are a pair of **coupled linear differential equations in time and position** for $V(x, t)$ and $I(x, t)$

Transmission Line Equation

We need to solve the **telegrapher's equations**,

Equation:

$$\frac{\partial V(x, t)}{\partial x} = - \left(L \frac{\partial I(x, t)}{\partial t} \right)$$

Equation:

$$\frac{\partial I(x, t)}{\partial x} = - \left(C \frac{\partial V(x, t)}{\partial t} \right)$$

The way we will proceed to a solution, and the way you always proceed when confronted with a pair of equations such as these, is to take a spatial derivative of one equation, and then substitute the second equation in for the spatial derivative in the first and you end up with...well, let's try it and see.

Taking a derivative with respect to x of [\[link\]](#)

Equation:

$$\frac{\partial^2 V(x, t)}{\partial x^2} = - \left(L \frac{\partial^2 I(x, t)}{\partial t \partial x} \right)$$

Now we substitute in for $\frac{\partial I(x, t)}{\partial x}$ from [\[link\]](#)

Equation:

$$\frac{\partial^2 V(x, t)}{\partial x^2} = LC \frac{\partial^2 V(x, t)}{\partial t^2}$$

It should be **very** easy for you to derive

Equation:

$$\frac{\partial^2 I(x, t)}{\partial \text{msup}} = LC \frac{\partial^2 I(x, t)}{\partial \text{msup}}$$

Oh, I know you all **love** differential equations! Well, let's take a look at these and just **think** for a minute. For **either** $V(x, t)$ or $I(x, t)$, we need to find a function that has some rather stringent requirements. First of all, the function must be of the form such that no matter whether we take its second derivative in space (x) or in time (t), it must end up differing in the way it behaves in x or t by no more than just a constant (LC).

In fact, we can be more specific than that. First $V(x, t)$ must have the same functional form for **both** its x and t variation. At most, the two derivatives must differ only by a constant. Let's try a "lucky" guess and let:

Equation:

$$V(x, t) = V_0 f(x - vt)$$

where V_0 is the amplitude of the voltage, and f is some function, of a form yet undetermined. Well

Equation:

$$\frac{\partial f(x - vt)}{\partial t} = - (vf')$$

and

Equation:

$$\frac{\partial^2 f(x - vt)}{\partial \text{msup}} = v^2 \frac{d f}{d}$$

Note also, that

Equation:

$$\frac{\partial^2 f(x - vt)}{\partial x^2} = f''$$

Now, let's take [\[link\]](#), [\[link\]](#), and [\[link\]](#) and substitute them into [\[link\]](#):
Equation:

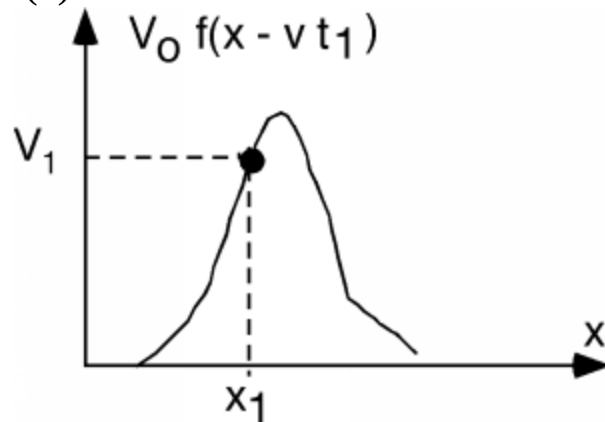
$$V_0 \frac{d f}{d x} = LC V_0 v^2 \frac{d f}{d x}$$

Our "lucky" guess works as a solution as long as
Equation:

$$v = \pm \frac{1}{\sqrt{LC}}$$

So, what is this $f(x - vt)$? We don't know yet what its actual functional form will be, but suppose at some point in time, t_1 , the function looks like [\[link\]](#).

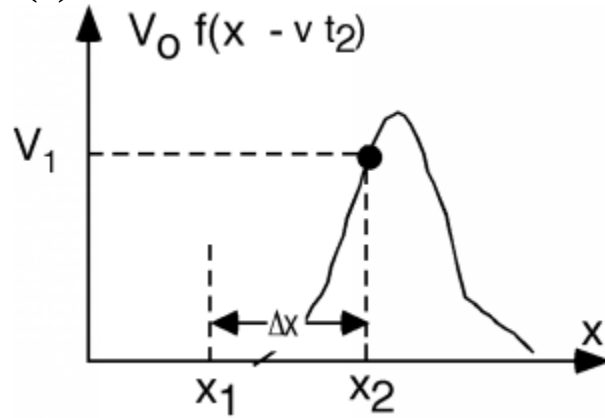
$f(x)$ At Some Point In Time



$f(x)$ at time t_1 .

At point x_1 , the function takes on the value V_1 . Now, let's advance to time t_2 . We look at the function and we see [\[link\]](#).

$f(x)$ At a Later Point In Time



$f(x)$ at a later t_2 .

If t increases from t_1 to t_2 then x will have to increase from x_1 to x_2 in order for the argument of f to have the same value, V_1 . Thus we find

Equation:

$$x_1 - vt_1 = x_2 - vt_2$$

which can be re-written as

Equation:

$$\frac{x_2 - x_1}{t_2 - t_1} = \frac{\Delta(x)}{\Delta(t)} \equiv v_p = \frac{1}{\sqrt{LC}}$$

where v_p is the velocity with which the function is moving along the x-axis! (We use the subscript "p" to indicated that what we have here is what is called the **phase velocity**. We will encounter another velocity called the **group velocity** a little later in the course.)

If we had "guessed" an $f(x + vt)$ for our function, it should be pretty easy to see that this would have given us a signal moving in the **minus** x direction, instead of the plus x direction. Thus we shall denote

Equation:

$$V_{\text{plus}} = V^+ f\left(x - \frac{1}{\sqrt{LC}}t\right)$$

the **positive** going voltage function and

Equation:

$$V_{\text{minus}} = V^- f\left(x + \frac{1}{\sqrt{LC}}t\right)$$

which is the negative going voltage function. Notice that since we are taking the **second** derivative of f with respect to t , we are free to choose either a $\frac{1}{\sqrt{LC}}$ or a $-\frac{1}{\sqrt{LC}}$ in front of the time argument inside f . Also note that these are our **only** choices for a solution. As we know from Differential Equations, a second order equation has, at most, two independent solutions.

Since $I(x, t)$ has the **same** differential equation describing its behavior, the solutions for I must also be of the exact same form. Thus we can let

Equation:

$$I_{\text{plus}} = I^+ f\left(x - \frac{1}{\sqrt{LC}}t\right)$$

represent the current function which goes in the positive x direction, and

Equation:

$$I_{\text{minus}} = I^- f\left(x + \frac{1}{\sqrt{LC}}t\right)$$

represent the negative going current function.

Now, let's take [\[link\]](#) and [\[link\]](#) and substitute them into [\[link\]](#):

Equation:

$$\frac{V^+}{\sqrt{LC}} f\left(x - \frac{1}{\sqrt{LC}}t\right) = LI^+ f\left(x - \frac{1}{\sqrt{LC}}t\right)$$

This can be solved for V^+ in terms of I^+ .

Equation:

$$V^+ = \sqrt{\frac{L}{C}} I^+ \equiv Z_0 I^+$$

where $Z_0 = \sqrt{\frac{L}{C}}$ is called the **characteristic impedance** of the transmission line. We will leave it as an exercise to the reader to ensure that indeed $\sqrt{\frac{L}{C}}$ has units of Ohms. For practice, and understanding about just how these equations work, the reader should ensure him/her self that

Equation:

$$V^- = - \left(\sqrt{\frac{L}{C}} I^- \right) \equiv - (Z_0 I^-)$$

Note the "subtle" difference here, with a "-" sign in front of the RHS of the equation!

We've been through lots of equations recently, so it is probably worth our while to summarize what we know so far.

1. The telegrapher's equations allow two solutions for the voltage and current on a transmission line. One moves in the x direction and the other moves in the $-x$ direction.
2. Both signals move at a constant velocity v_p given by [\[link\]](#).

3. The voltage and current signals are related to one another by the characteristic impedance Z_0 , with [\[link\]](#)

Equation:

$$v_p = \frac{1}{\sqrt{LC}}$$

Equation:

$$Z_0 = \sqrt{\frac{L}{C}}$$

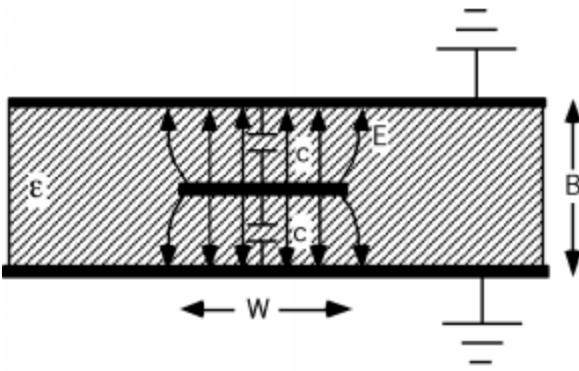
$$\frac{V^+}{I^+} = Z_0$$

$$\frac{V^-}{I^-} = -Z_0$$

Transmission Line Examples

As an example, and also because it even has some practical importance, let's look at one kind of transmission line. It is called a **stripline** and it looks like [\[link\]](#). It consists of a flat conductor, located between two ground planes. It is supported by an insulating dielectric with dielectric constant ϵ . This is kind of like the situation you would find on a multi-level PC board, where perhaps the bus lines would be running on an inner layer with ground planes above and below them.

A Stripline



Between the center conductor and the ground plane, there will be some capacitance, C . If we can assume that the electric field is more or less confined to the regions between the strip conductor and the ground plane (which occurs when the ratio of $\frac{W}{B}$ is not too small) then for either capacitor (assuming unit length into the picture) we will get a value

Equation:

$$C = \frac{\epsilon W}{\frac{B}{2}}$$

since the value of a capacitor is just the dielectric constant times the area of the plates, divided by the spacing of the plates.

Looking quickly at [\[link\]](#) you might think the two capacitors are in series, but you would be wrong! Note that each capacitor has one lead connected

to the center conductor and the other lead connected to ground, and so the two capacitors are in fact, in parallel, and hence their capacitances add. Thus, for the capacitance per unit length for this line, we can write:

Equation:

$$= \frac{4\epsilon W}{B}$$

It can be shown (although we won't do it here) that for **any** transmission line where the electric and magnetic fields are perpendicular to one another (called **TEM** or **transverse electromagnetic**) the speed of propagation of the wave down the line is just

Equation:

$$\begin{aligned} v_p &= \frac{c}{\frac{\epsilon}{\epsilon_0}} \\ &= \frac{3 \times 10^8 \frac{m}{s}}{\sqrt{\epsilon_r}} \end{aligned}$$

Where ϵ_r is called the **relative dielectric constant** for the material. Well, we also know that

Equation:

$$v_p = \frac{1}{\sqrt{\quad}}$$

From which we can write

Equation:

$$\begin{aligned} &= \frac{1}{v_p^2} \\ &= \frac{B}{v_p^2 4\epsilon W} \end{aligned}$$

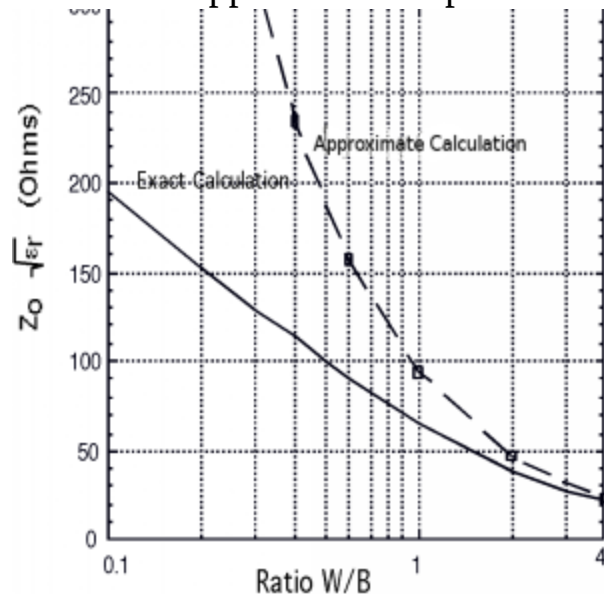
We can now insert this value for \quad into the expression for Z_0 , the impedance of the line.

Equation:

$$\begin{aligned}
 Z_0 &= \frac{1}{\frac{v_p^2 4\epsilon W}{B}} \\
 &= \frac{B}{4\epsilon W v_p} \\
 &= \frac{B}{4\epsilon W \frac{c}{\sqrt{\epsilon_r}}}
 \end{aligned}$$

And so, we have derived an equation for the impedance Z_0 of the line in terms of the dimensions W and B , the dielectric constant of the insulating material, ϵ , and c , the speed of light. How good is this expression, and in particular how good is our assumption that the electric field is all confined to the region under the conductor? Not so great actually [\[link\]](#).

Exact and Approximate Impedance For a Stripline



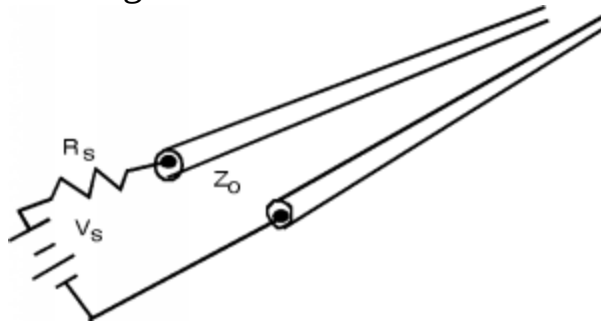
Exact and approximate Z_0 for a stripline

[\[link\]](#) shows the results from using [\[link\]](#) and a more exact calculation, which takes into account the fringing fields. As you can see we have to get the ratio $\frac{W}{B}$ up to about 4 or so before the two match. But at least we get the right behavior and we're not totally out of the ball park.

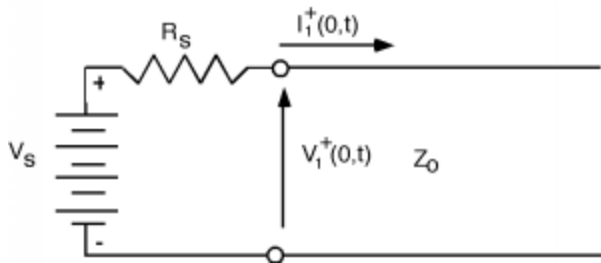
Exciting a Line

We will now go on and look at what happens when we excite the line. Let's take a DC voltage source with a source internal impedance R_s and connect it to our semi-infinite line. The sketch in [\[link\]](#) is sort of awkward looking, and will be hard to analyze, so let's make a more "schematic like" drawing [\[link\]](#), keeping in mind that it is a situation such as [\[link\]](#) which we trying to represent.

Exciting a Transmission Line



Schematic Representation



Why have we shown an I^+ and a V^+ but not V^- or I^- ? The answer is, that if the line is semi-infinite, then the "other" end is at infinity, and we know there are no sources at infinity. The current flowing through the source resistor is just I_1^+ , so we can do a KVL around the loop

Equation:

$$V_s - I_1^+(0, t)R_s - V_1^+(0, t) = 0$$

Substituting for I_1^+ in terms of V_1^+ using [this equation](#):

Equation:

$$V_s - \frac{V_1^+(0, t)}{Z_0} R_s - V_1^+(0, t) = 0$$

Which we re-write as

Equation:

$$V_1^+(0, t) \left(1 + \frac{R_s}{Z_0} \right) = V_s$$

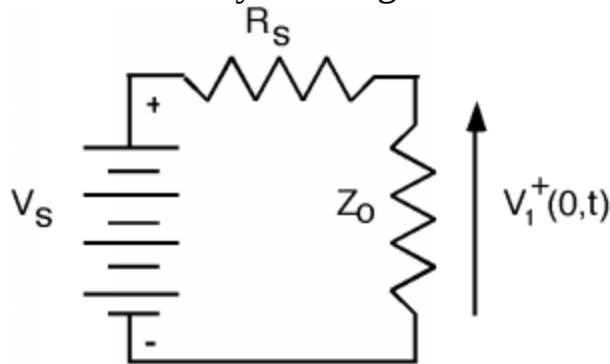
Or, on solving for $V_1^+(0, t)$:

Equation:

$$V_1^+(0, t) = \frac{Z_0}{Z_0 + R_s} V_s$$

This **should** look both reasonable and familiar to you. The line and the source resistance are acting as a [voltage divider](#). In fact, [\[link\]](#) is just the usual voltage divider equation for two resistors in series. Thus, the generator can not tell the difference between a semi-infinite transmission line of characteristic impedance Z_0 and a resistor with a resistance of the same value [\[link\]](#).

Line is Initially a Voltage Divider!

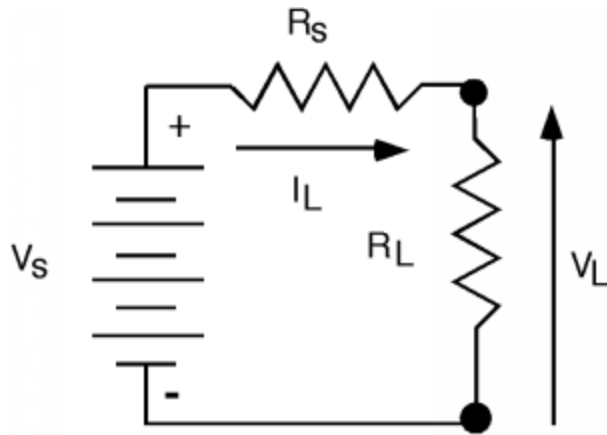


Have you ever heard of "300 Ω twin-lead" or maybe "75 Ω co-ax" and wondered why people would want to use wires with such a high resistance value to bring a TV signal to their set? Now you know. The 300 Ω

characterization is not a measure of the resistance of the wire, rather it is a specification of the transmission line's impedance. Thus, if a TV signal coming from your antenna has a value of, say, $30\mu\text{V}$, and it is being brought down from the roof with 300Ω twin-lead, then the current flowing in the wires is $I = \frac{30\mu\text{V}}{300\Omega} = 100\text{ nA}$, which is a very small current indeed!

Why then, did people decide on 300Ω ? An antenna which is just a half-wavelength long (Which turns out to be both a convenient and efficient choice for signals in the 100 MHz ($\lambda \simeq 3\text{m}$) range) acts like a voltage source with a source resistance of about 300Ω . If you remember from ELEC 242, when we have a source with a source resistance R_s and a load resistor with load resistance value R_L [\[link\]](#), you calculate the power delivered to the load using the following method.

Power Transfer To a Load



P_L , the power in the load, is just product of the voltage across the load times the current through the load. We can use the voltage divider law to find the voltage across R_L and the resistor sum law to find the current through it.

Equation:

$$\begin{aligned}
 P_L &= V_L I_L \\
 &= \frac{R_L}{R_L + R_s} V_s \frac{V_s}{R_L + R_s} \\
 &= \frac{R_L}{(R_L + R_s)^2} V_s^2
 \end{aligned}$$

If we take the derivative of [\[link\]](#) with respect to R_L , the load resistor (which we assume we can pick, given some predetermined R_s) we have (ignoring the V_s^2),

Equation:

$$\begin{aligned}\frac{d}{dR_L}(P_L) &= \frac{1}{(R_L+R_s)^2} \frac{2R_L}{(R_L+R_s)^3} \\ &= 0\end{aligned}$$

Putting everything on $(R_L + R_s)^3$ and then just looking at the numerator:

Equation:

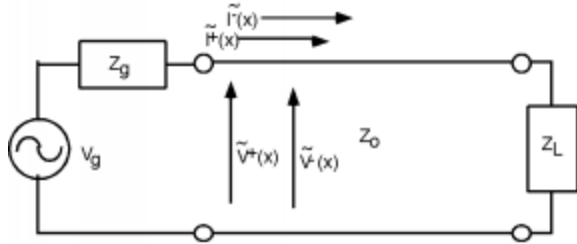
$$R_L + R_s - 2R_L = 0$$

Which obviously says that for maximum power transfer, you want your load resistor R_L to have the same value as your source resistor R_s ! Thus, people came up with 300Ω twin lead so that they could maximize the energy transfer between the TV antenna and the transmission line bringing the signal to the TV receiver set. It turns out that for a co-axial transmission line (such as your TV cable) 75Ω minimizes the signal loss, which is why that value was chosen for CATV.

Terminated Lines

If, on the other hand, we have a finite line, terminated with some load impedance, we have a somewhat more complicated problem to deal with [\[link\]](#).

A Finite Terminated Transmission Line



There are several things we should note **before** we head off into equation-land again. First of all, unlike the transient problems we looked at in a [previous chapter](#), there can be no more than **two** voltage and current signals on the line, just V^+ and V^- , (and I^+ and I^-). We no longer have the luxury of having V_1^+ , V_2^+ , etc., because we are talking now about a **steady state system**. All of the transient solutions which built up when the generator was first connected to the line have been summed together into just two waves.

Thus, on the line we have a single **total voltage function**, which is just the sum of the positive and negative going voltage waves

Equation:

$$V(x) = V^+ e^{-i\beta x} + V^- e^{i\beta x}$$

and a total current function

Equation:

$$I(x) = I^+ e^{-i\beta x} + I^- e^{i\beta x}$$

Note also that until we have solved for V^+ and V^- , we do not know $V(x)$ or $I(x)$ anywhere on the line. In particular, we do not know $V(0)$ and $I(0)$

which would tell us what the apparent impedance is looking into the line.

Equation:

$$\begin{aligned} Z_{\text{in}} &= Z(0) \\ &= \frac{V^+ + V^-}{I^+ + I^-} \end{aligned}$$

Until we know what kind of impedance the generator is seeing, we can not figure out how much of the generator's voltage will be coupled to the line! The input impedance looking into the line is now a function of the load impedance, the length of the line, and the phase velocity on the line. We have to solve this **before** we can figure out how the line and generator will interact.

The approach we shall have to take is the following. We will start at the **load** end of the line, and in a manner similar to the one we used previously, find a relationship between V^+ and V^- , leaving their actual magnitude and phase as something to be determined later. We can then propagate the two voltages (and currents) back down to the input, determine what the input impedance is by finding the ratio of $(V^+ + V^-)$ to $(I^+ + I^-)$, and from this, and knowledge of properties of the generator and its impedance, determine what the actual voltages and currents are.

Let's take a look at the load. We again use KVL and KCL ([\[link\]](#)) to match voltages and currents in the line and voltages and currents in the load:

Equation:

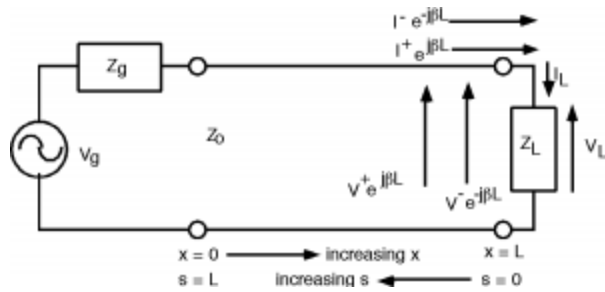
$$V^+ e^{-(i\beta L)} + V^- e^{i\beta L} = V_L$$

and

Equation:

$$I^+ e^{-(i\beta L)} + I^- e^{i\beta L} = I_L$$

Doing Kirchoff at the End of the Line



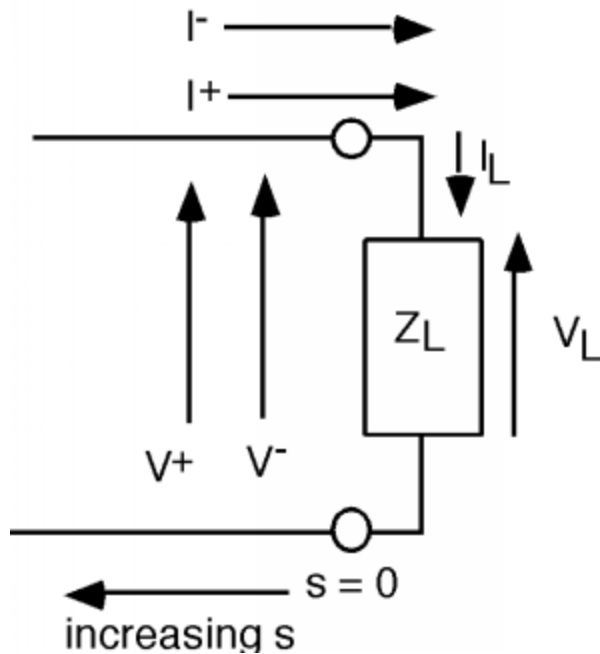
Change variables!

Now, we **could** substitute $\frac{\pm(V)}{Z_0}$ for the two currents on the line and $\frac{V_L}{Z_L}$ for I_L , and then try to solve for V^- in terms of V^+ using [\[link\]](#) and [\[link\]](#) but we can be a little clever at the outset, and make our (complex) algebra a good bit cleaner [\[link\]](#). Let's make a change of variable and let

Equation:

$$s \equiv L - x$$

$s=0$ at the Load and So the Exponentials Go Away!



This then gives us for the voltage on the line (using $x = L - s$)

Equation:

$$V(s) = V^+ e^{-(i\beta L)} e^{i\beta L} + V^- e^{i\beta L} e^{-(i\beta L)}$$

Usually, we just fold the (constant) phase terms $e^{\pm(i\beta L)}$ terms in with the V^+ and V^- and so we have:

Equation:

$$V(s) = V^+ e^{i\beta s} + V^- e^{-(i\beta s)}$$

Note that when we do this, we now have a **positive** exponential in the first term associated with V^+ and a **negative** exponential associated with the V^- term. Of course, we also get for $I(s)$:

Equation:

$$I(s) = I^+ e^{i\beta s} + I^- e^{-(i\beta s)}$$

This change now moves our origin to the **load** end of the line, and reverses the direction of positive motion. **But**, now when we plug into $e^{i\beta s}$ the value for s at the load ($s = 0$), the equations simplify to:

Equation:

$$V^+ + V^- = V_L$$

and

Equation:

$$I^+ + I^- = I_L$$

which we then re-write as

Equation:

$$\frac{V^+}{Z_0} - \frac{V^-}{Z_0} = \frac{V_L}{Z_L}$$

This is beginning to look almost exactly like a [previous chapter](#). As a reminder, we solve [\[link\]](#) for V_L

Equation:

$$V_L = \frac{Z_L}{Z_0} V^+ - V^-$$

and substitute for V_L in [\[link\]](#)

Equation:

$$V^+ + V^- = \frac{Z_L}{Z_0} V^+ - V^-$$

From which we then solve for the reflection coefficient Γ_ν , the ratio of V^- to V^+ .

Equation:

$$\frac{V^-}{V^+} \equiv \Gamma_\nu = \frac{Z_L - Z_0}{Z_L + Z_0}$$

Note that since, in general, Z_L will be complex, we can expect that Γ_ν will also be a complex number with both a magnitude $|\Gamma_\nu|$ and a phase angle θ_Γ . Also, as with the case when we were looking at transients, $|\Gamma_\nu| < 1$.

Since we now know V^- in terms of V^+ , we can now write an expression for $V(s)$ the voltage anywhere on the line.

Equation:

$$V(s) = V^+ e^{i\beta s} + V^- e^{-(i\beta s)}$$

Note again the change in signs in the two exponentials. Since our spatial variable s is going in the opposite direction from x , the V^+ phasor now goes as $i\beta s$ and the V^- phasor now goes as $-(i\beta s)$.

We now substitute in $\Gamma_\nu V^+$ for V^- in [\[link\]](#), and for reasons that will become apparent soon, factor out an $e^{i\beta s}$.

Equation:

$$\begin{aligned} V(s) &= V^+ e^{i\beta s} + \Gamma_\nu V^+ e^{-(i\beta s)} \\ &= V^+ e^{i\beta s} + \Gamma_\nu e^{-(i\beta s)} \\ &= V^+ e^{i\beta s} [1 + \Gamma_\nu e^{-(2i\beta s)}] \end{aligned}$$

We could have also written down an equation for $I(s)$, the current along the line. It will be a good test of your understanding of the basic equations we are developing here to show yourself that indeed

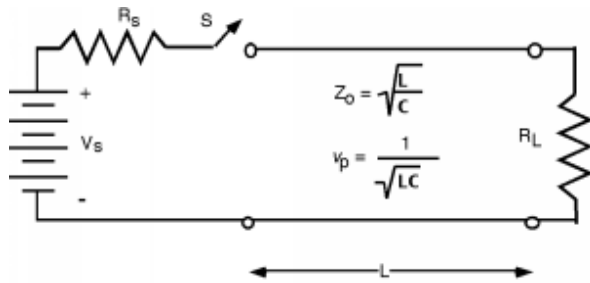
Equation:

$$I(s) = \frac{V^+ e^{i\beta s}}{Z_0} [1 - \Gamma_\nu e^{-(2i\beta s)}]$$

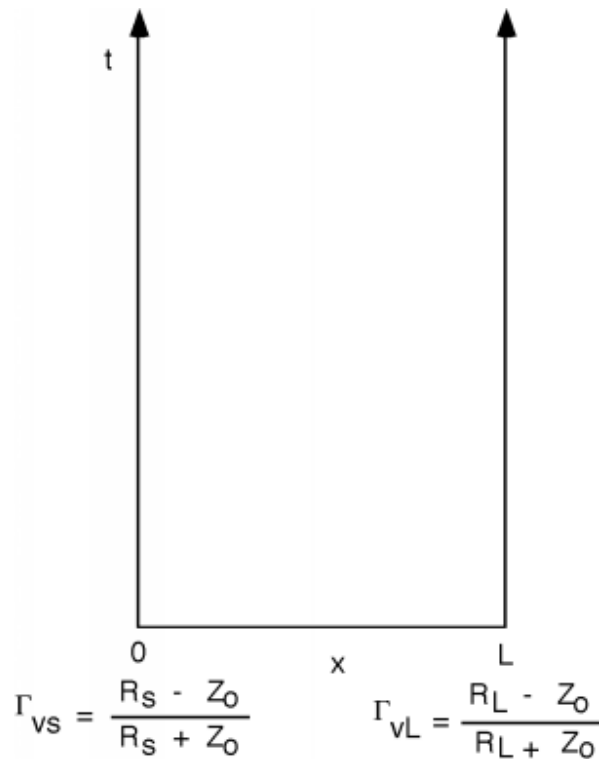
Bounce Diagrams

Now this **new** V_2^+ will head back towards the load and ...Hmmm... things are going to get kind of messy and complicated. Fortunately for us, transmission line engineers came up with a scheme for keeping track of all of the waves bouncing back and forth on the line. The scheme is called a **bounce diagram**. A bounce diagram consists of a horizontal distance line, which represents distance along the transmission line, and a vertical time axis, which represents time since the battery was first connected to the line. Just to keep things conceptually clear, we usually first start out by showing the line, the battery, the load and a switch, S, which is used to connect the source to the line. It doesn't hurt to make a little sketch like [\[link\]](#), and write down the length of the line, Z_0 and v_p , along with the source and load resistances. Now we draw the bounce diagram, which is shown in [\[link\]](#)

Transient Problem

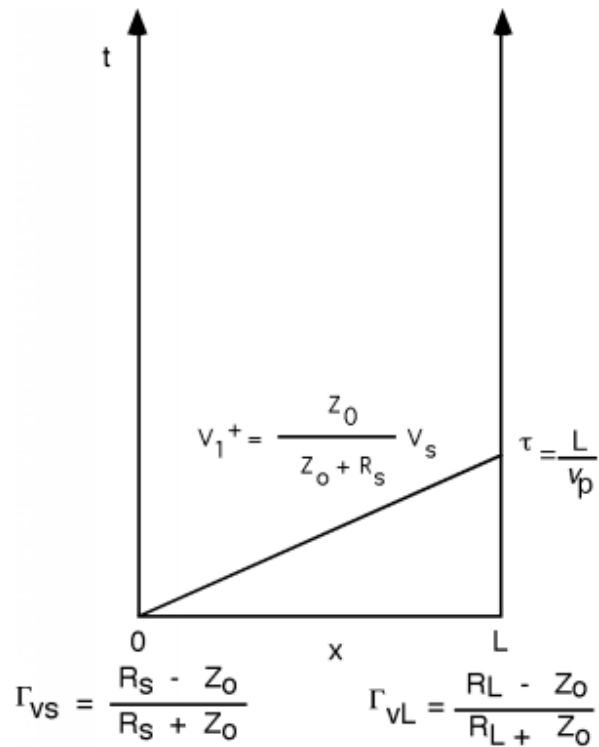


A "Bounce Diagram"



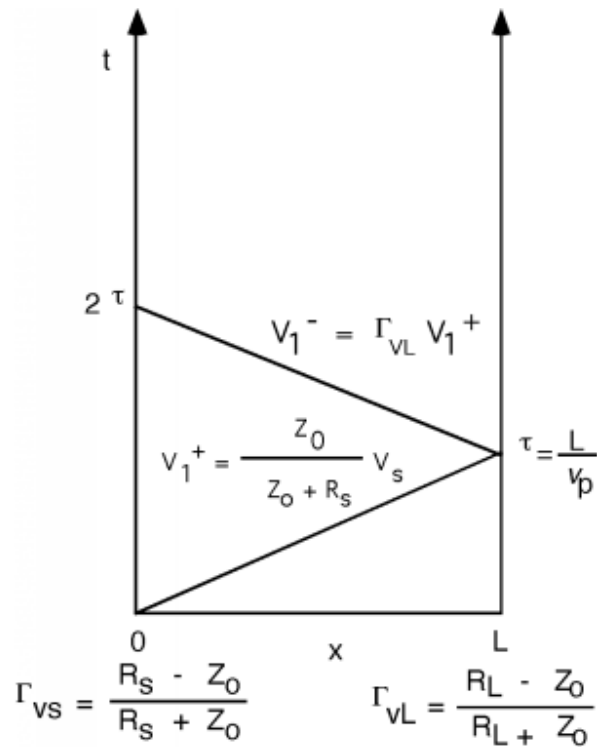
Normally, you would not put the formula for Γ_{vS} and Γ_{vL} by 0 and L in the diagram, but rather their values. This will become clear when we do an example. The next thing we do is calculate V_1^+ and draw a straight line on the bounce diagram (nominally at a slope of $\frac{1}{v_p}$) which will represent the initial signal going down the line. We mark a $\tau = \frac{L}{v_p}$ on the vertical axis to show how long it takes for the wave to reach the end of the line [\[link\]](#).

Diagram With First Wave



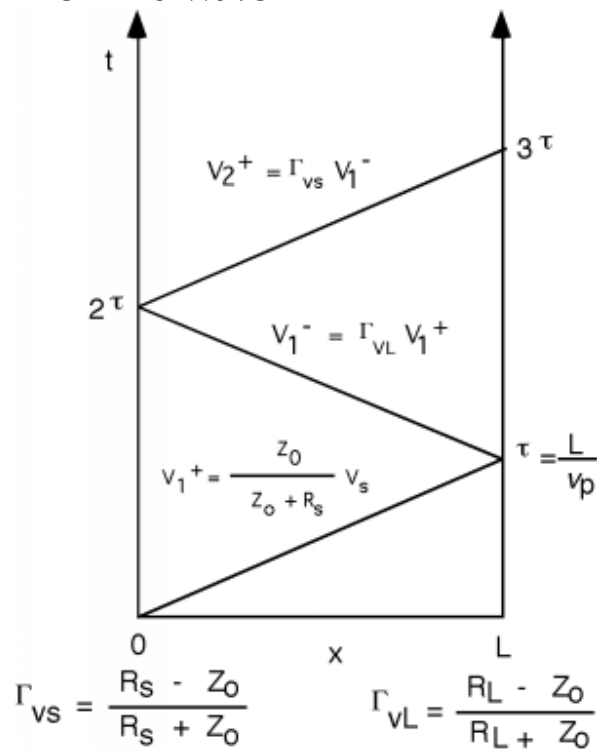
Once the initial wave hits the load, a second, reflected wave $V_1^- = \Gamma_{vL} V_1^+$ is sent back the other way. So we add it to the bounce diagram. This is shown in [\[link\]](#). Since all of the waves move with the same phase velocity, we should be careful to draw all of the lines with the same slope. Note that the time when the reflected wave hits the generator end is a total round trip time of 2τ . (This simple concept is one which students often forget come test time, so be forewarned!)

Adding the First Reflected Wave



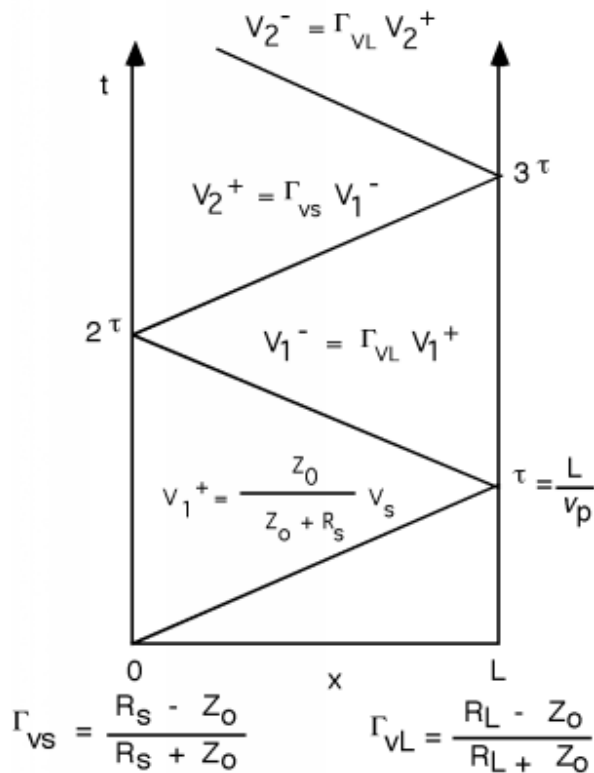
We saw that the next thing that happens is that another wave is reflected from the generator, so we add that to the bounce diagram as well. This is shown in [\[link\]](#).

The Third Wave



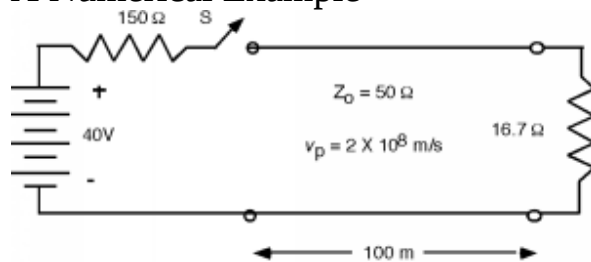
Finally, one last wave, as we are almost bounced right off the diagram, as shown in [\[link\]](#)!

And the Fourth



OK, so we've got a bounce diagram, so what? Having the diagram is only part of the solution. We still have to see what good they are. Let's do a numerical example, as it is maybe a little more illustrative, and certainly will be easier to write out than all these ratios all the time. We will just pick some typical numbers, and then work out the answers. Let's let $V_S = 40(V)$, $R_S = 150(\Omega)$, $Z_0 = 50(\Omega)$ and $R_L = 16.7(\Omega)$. The line will be 100m long, and $v_p = 2 \times 10^8 \frac{m}{s}$ [\[link\]](#).

A Numerical Example



First we calculate the reflection coefficients

Equation:

$$\begin{aligned}
 \Gamma_{vL} &= \frac{R_L - Z_0}{R_L + Z_0} \\
 &= \frac{16.7 - 50}{16.7 + 50} \\
 &= -0.50
 \end{aligned}$$

and

Equation:

$$\begin{aligned}
 \Gamma_{vS} &= \frac{R_S - Z_0}{R_S + Z_0} \\
 &= \frac{150 - 50}{150 + 50} \\
 &= 0.50
 \end{aligned}$$

The initial voltage signal V_1^+ is

Equation:

$$\begin{aligned}
 V_1^+ &= \frac{50}{50 + 150} 40 \\
 &= 10(V)
 \end{aligned}$$

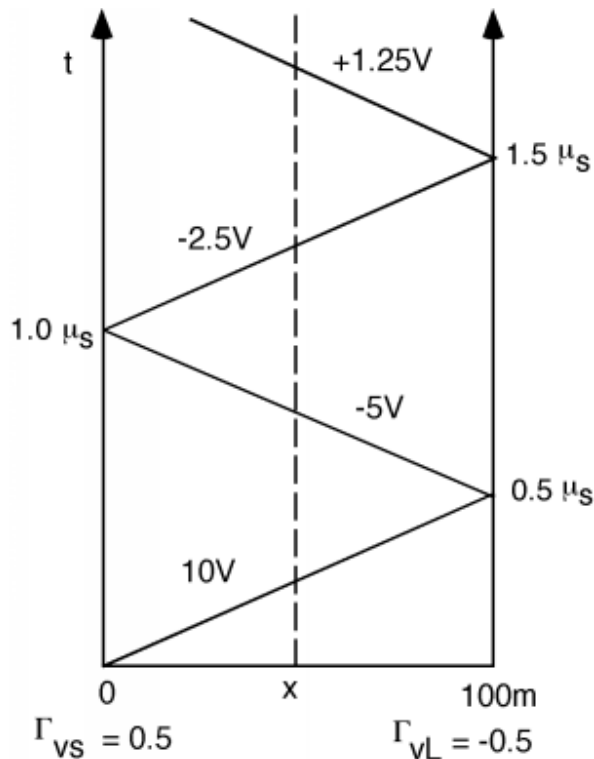
and the propagation time is

Equation:

$$\begin{aligned}
 \tau &= \frac{L}{v_p} \\
 &= \frac{100(m)}{(2 \times 10^8) \frac{m}{s}} \\
 &= 0.5(\mu, s)
 \end{aligned}$$

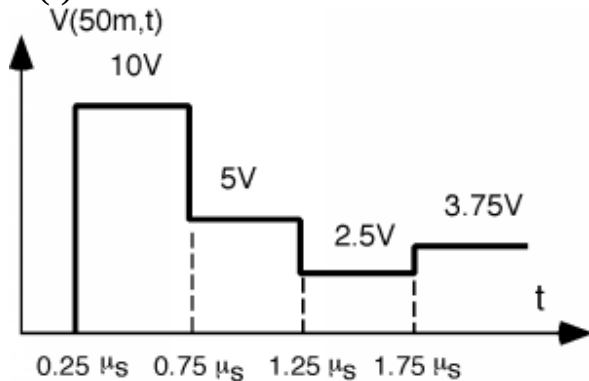
So we draw the bounce diagrams seen in [\[link\]](#).

The Bounce Diagram



Now, here's how we use a bounce diagram, once we have it. Suppose we want to know what $V(t)$, the voltage as a function of time, would look like half-way down the line. We draw a vertical line at the place we are interested in (the dotted line in [\[link\]](#)) and then just go up along the line, adding voltage to whatever we had before whenever we cross one of the "bouncing" signal lines. Thus for the line as shown we would have for $V(t)$ what we see in [\[link\]](#).

$V(t)$ at 50m Down the Line



For the first 0.25 μs we have no voltage, because V_1^+ has not reached the half-way point yet. The voltage then jumps to +10V when V_1^+ comes by. It stays

like that until the $-5V V_1^-$ comes by $0.5\mu s$ later. The voltage then remains constant at $5V$ until the $-2.5V V_2^+$ comes along to drop the total voltage down to only 2.5 volts. When V_2^- comes along, it has been switched back to a positive voltage wave by the negative load reflection coefficient, and so now the voltage jumps back up to $3.75V$. It will keep oscillating back and forth until it finally settles down to some asymptotic value.

What will that asymptotic value be? One approach is to write down the following equation.

Equation:

$$V(x, \infty) = V_1^+ (1 + \Gamma_L + \Gamma_L \Gamma_S + \Gamma_L^2 \Gamma_S + \dots)$$

Which we can re-write as

Equation:

$$V_1^+ (1 + \Gamma_L \Gamma_S + (\Gamma_L \Gamma_S)^2 + \dots) + \Gamma_L V_1^+ (1 + \Gamma_L \Gamma_S + (\Gamma_L \Gamma_S)^2 + \dots)$$

Now, remembering the infinite sum relationship:

Equation:

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

for $|x| < 1$ (which is **always** the case for a reflection coefficient). We can substitute [\[link\]](#) for the terms inside the parentheses in [\[link\]](#) and we get

Equation:

$$\begin{aligned} V(x, \infty) &= V_1^+ \left(\frac{1}{1-\Gamma_L \Gamma_S} + \frac{\Gamma_L}{1-\Gamma_L \Gamma_S} \right) \\ &= V_1^+ \frac{1+\Gamma_L}{1-\Gamma_L \Gamma_S} \end{aligned}$$

We will leave it as an exercise to the reader to show that if we substitute [\[link\]](#), [\[link\]](#) and finally [\[link\]](#) into [\[link\]](#) we will eventually get:

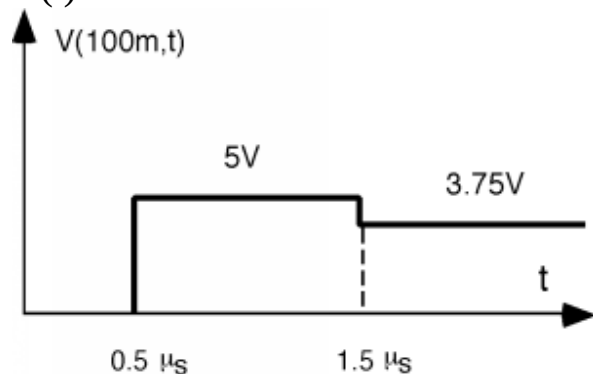
Equation:

$$V(x, \infty) = \frac{R_L}{R_L + R_S} V_S$$

Look back at [\[link\]](#) and see if [\[link\]](#) makes any sense. It should. If we wait long enough, it is reasonable to expect that any "transmission line" effects should go away, and we would be back to the same situation we would have if the line was just some wire connecting the source to the load. In this case, the load resistor and the source resistor would form a voltage divider, and we would expect the voltage across the load to be determined by the voltage divider equation. That's all [\[link\]](#) is saying!

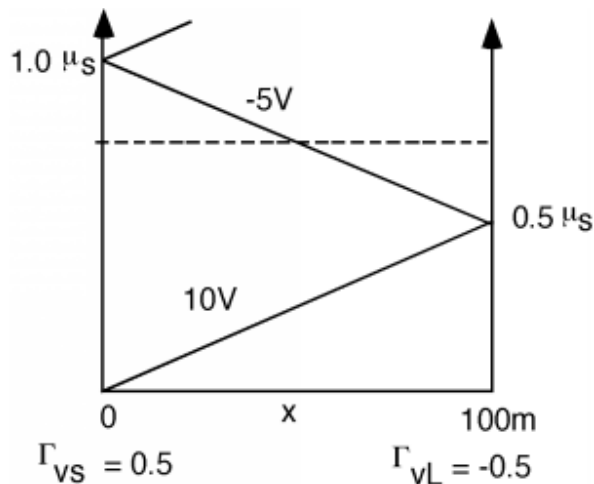
What do we do if we want, say, the voltage across the load with time? To do this we move up the RHS of the bounce diagram, and count voltage waves as we move across them. We start out at zero, of course, and do not see anything until we get to 0.5ms. Then we cross the 10V V_1^+ wave **and** we cross the -5V V_1^- wave at the same time, so the voltage only goes up to +5V. Likewise, another 1ms later, we cross both the -2.5V V_2^+ **and** the +1.25V V_2^- wave, and so the voltage ends up at the 3.75V position [\[link\]](#).

V(t) Across the Load



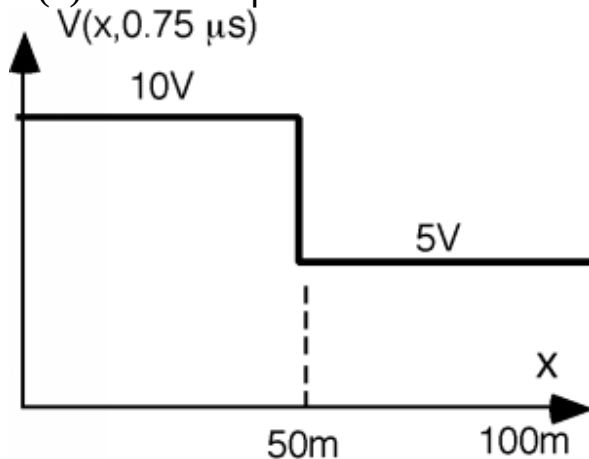
We can also use the bounce diagram to find the voltage as a function of position, for some fixed time, t_0 [\[link\]](#).

Finding V(x) at t=0.75μs



To do this, we draw a horizontal line at the time we are interested in, say $0.75\mu\text{s}$. Now, for each position x , we go from the bottom of the diagram, up to the horizontal line, adding up voltage as we go. Thus for the example: we get what we see in [\[link\]](#). For the first half of the line, we cross the $+10\text{V } V_1^+$, but that's it. For the second half of the line we cross **both** the $+10\text{V}$ line as well as $-5\text{V } V_1^-$ wave, and so the voltage drops down to 5V .

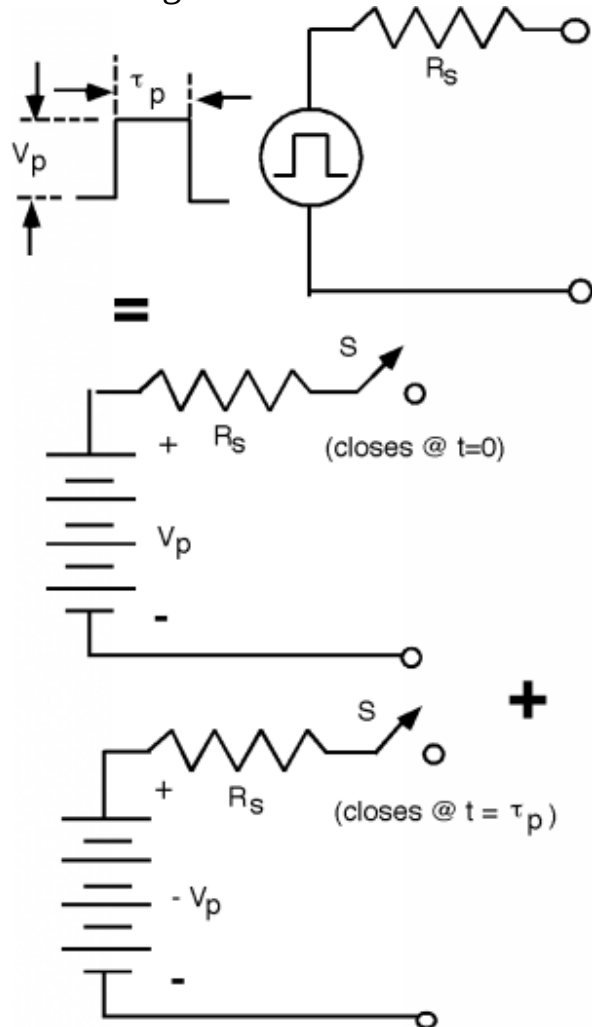
$V(x)$ at $t = 0.75\mu\text{s}$



Of particular interest to many of you will be the way in which a **pulse** moves down a line and is reflected etc. This is also quite easy to do with a reflection diagram, if we simply break the pulse into two waves, one which has a **positive** swing at $t = 0$ and another which is a **negative** going wave at $t = \tau_p$, where τ_p is the pulse width of the pulse being generated. The way we do this is suggested in [\[link\]](#). We replace the pulse generator with two battery/switch combinations. The first circuit is just like we have seen so far, with a battery equal to the open circuit pulse height of the generator, and a switch which

closes at $t = 0$. The second circuit has a battery with an amplitude of **minus** the pulse height, and a switch which closes at $t = \tau_p$, the pulse width of the pulse itself.

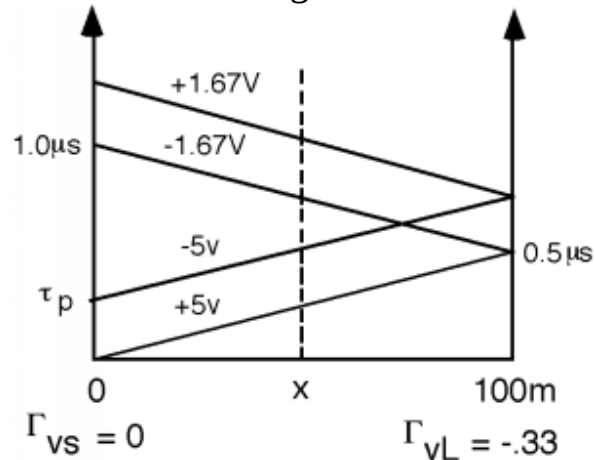
Simulating a Pulse With Two Batteries and Two Switches



By superposition, we can just add these two generators, one after the other, and see how the pulse goes down the line. Suppose V_p is 10 volts, $\tau_p = 0.25(\mu, s)$, $R_s = 50(\Omega)$, $Z_0 = 50(\Omega)$ and $R_L = 25(\Omega)$. With the numbers, we find that $V_1^+ = 25(V)$. $\Gamma_{vL} = \frac{-1}{3}$ and $\Gamma_{vS} = 0$. Let's assume that the propagation time on the line is still $0.5\mu s$ to get from one end of the line to the other.

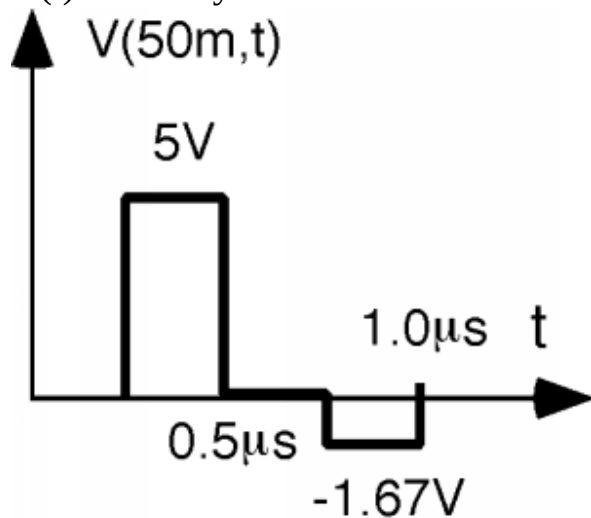
We draw the [bounce diagram](#), and launch **two waves**, one which leaves at $t = 0$ has an amplitude of $V_1^+ = 5(V)$. The second wave leaves at a time τ_p , later, and has an amplitude of $-5V$.

Pulse Bounce Diagram



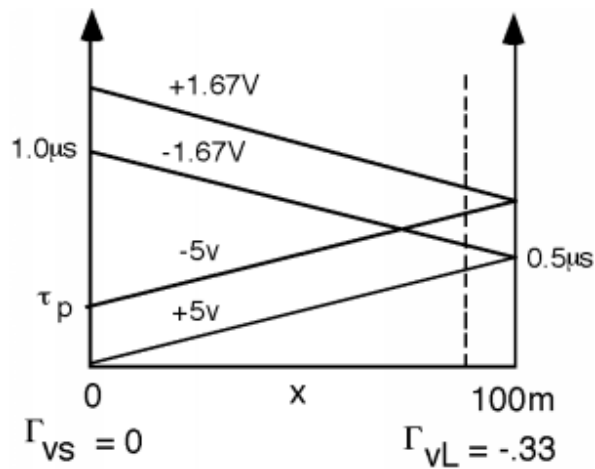
Now when we want to see what the voltage as a function of time looks like, we again draw a line up the middle, and add voltages as we cross them. Here we see, again, no voltage until we cross the first wave at $0.25\mu s$, which pops us up to $+5V$. At a time $0.25\mu s$ later however, the $-5V$ wave comes along, and we go back down to zero. At $t = 0.75(\mu, s)$, the reflected $-1.67V$ pulse comes along, and so we see that. Since the source is matched to the line, $\Gamma_{VS} = 0$ and so this is the end of the story [\[link\]](#).

$V(t)$ Half-way Down the Line



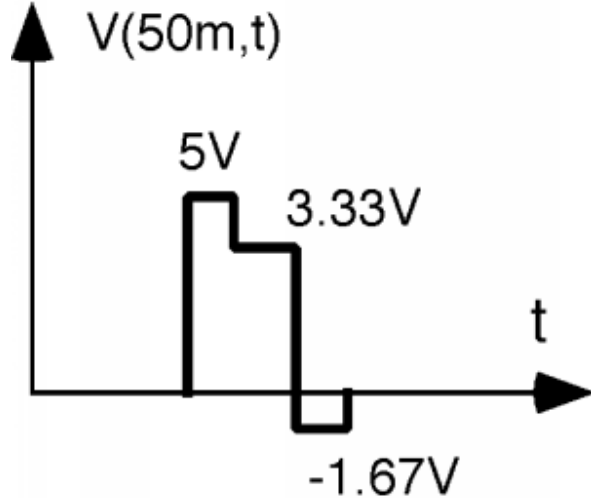
You can get somewhat more interesting waveforms if you go someplace where the two pulses at least partially overlap. Let's look at say, $x = 87.5(m)$. [Here](#) is the bounce diagram.

Finding $V(t)$ Near the Load



And [here](#) is the voltage waveform we get.

V(t) Near the Load



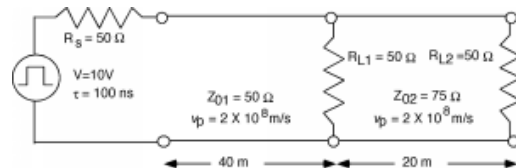
This time the 1.67V pulse gets to us before the +5V pulse has completely passed, and so we drop from 5V to 3.33V. Then, when the -5V wave goes by, we drop down to -1.67V for a little while, until the +1.67V wave comes along to bring us back to zero.

Cascaded Lines

We can use bounce diagrams to handle somewhat more complicated problems as well.

Arnold Aggie decides to add an additional ethernet interface to the one already connected to his computer. He decides just to add a "T" to the terminal where the cable is connected to his "thin-net" interface, and add on some more wire. Unfortunately, he is not careful about the coaxial cable he uses, and so he has some 75Ω TV co-ax instead of the 50Ω ethernet cable. He ends up with the situation shown [here](#). This kind of problem is called a **cascaded line problem** because we have two different lines, one hooked up after the other. The analysis is similar to what we have done before, just a little more complicated is all.

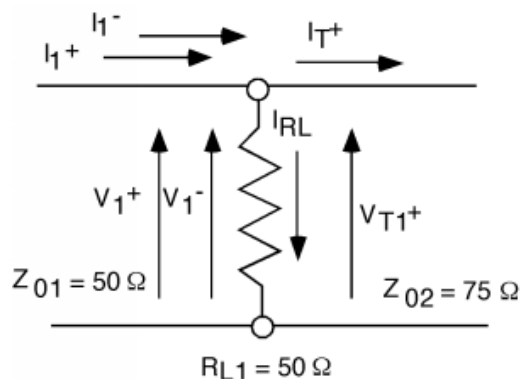
Cascaded Line Problem



We will have to do a little more thinking before we can draw out the bounce diagram for this problem. The driver for ethernet cable coming to Arnold's computer can be modeled as a 10V (open circuit) source with a 50Ω internal impedance. Since the source does not (initially) know anything about how the line it is driving is terminated, the first signal V_1^+ will be the same as in our initial problem, in this case just a +5V signal headed down the line.

Let's focus on the "T" for a minute [\[link\]](#).

At the Junction



V_1^+ is incident on the junction. When it hits the junction, there will be a reflected wave V_1^- and also now, a transmitted wave V_{T1}^+ . Since the incident wave can not tell the difference between a 75Ω resistor and a 75Ω transmission line, it **thinks** it is seeing a termination resistor equal to a 50Ω resistor (R_{L1}) in parallel with a 75Ω resistor (the second line). 50Ω in parallel with 75Ω is 30Ω . Let's call this "apparent" load resistor R'_L .

), so that we can then calculate $\Gamma_{V_{12}}$, the first voltage reflection coefficient in going from line 1 to line 2 as:

Equation:

$$\begin{aligned}\Gamma_{V_{12}} &= \frac{R'_L - Z_{01}}{R'_L + Z_{01}} \\ &= \frac{30 - 50}{30 + 50} \\ &= -0.25\end{aligned}$$

Note that we **could** have started from scratch and written down KVLs and KCLs for the junction

Equation:

$$V_1^+ + V_1^- = V_{T1}^+$$

and

Equation:

$$I_1^+ + I_1^- = I_{RL} + I_{T1}^+$$

Then, by re-writing [\[link\]](#) in terms of voltage and impedances we have:

Equation:

$$\frac{V_1^+}{Z_{01}} - \frac{V_1^-}{Z_{01}} = \frac{V_{T1}^+}{Z_{02}} + \frac{V_{T1}^+}{R_L}$$

We now have two equations with two unknowns (V_1^- and V_{T1}^+). By solving [\[link\]](#) for V_{T1}^+ and then plugging that into [\[link\]](#), we could get the ratio of V_1^- to V_1^+ , or the voltage reflection coefficient. The interested reader can confirm that indeed, you get the very same result this way.

In order to completely solve this problem, we also need to know V_{T1}^+ , the transmitted wave as well. Since [\[link\]](#) says V_{T1}^+ is just the sum of the incident and reflected waves on the first line

Equation:

$$V_{T1}^+ = V_1^+ + \Gamma_{V_{12}} V_1^+$$

We can thus write

Equation:

$$\frac{V_{T1}^+}{V_L^+} = 1 + \Gamma_{V_{12}} = \frac{R'_L + Z_{01}}{R'_L + Z_{01}} + \frac{R'_L - Z_{01}}{R'_L + Z_{01}} = \frac{2R'_L}{R'_L + Z_{01}} = \frac{60}{30 + 50} = 0.75 \equiv T_{V_{12}}$$

An important thing to note is that

Equation:

$$T_V = 1 + \Gamma_V$$

NOT

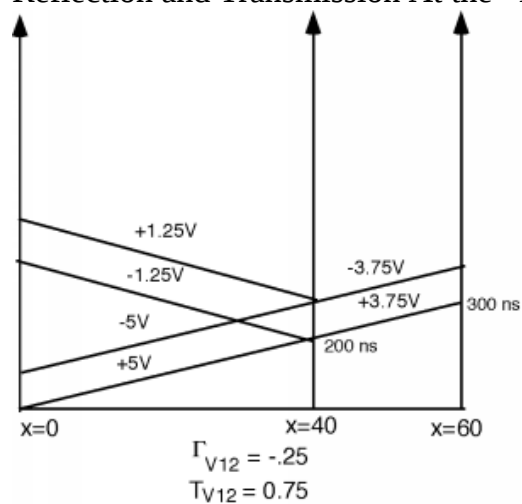
Equation:

$$T_V + \Gamma_V = 1$$

We do not "conserve" voltage at a termination, in the sense that the reflected and transmitted voltage have to add up to be the incident voltage. Rather, the transmitted voltage is the **sum** of the incident voltage and the reflected voltage, so that we can obey Kirchoff's voltage law.

We can now start to make our bounce diagram. We propagate a +5V wave and a -5V wave (separated by 100ns) down towards the junction. Since the line is 40m long, and the waves move at $2 \times 10^8 \frac{m}{s}$, it takes 200ns for them to get to the junction. There, a -1.25V wave is reflected back towards the source, and a +3.75V wave is transmitted into the second transmission line [\[link\]](#).

Reflection and Transmission At the "T"



Since the load for the second line is 50Ω , and the characteristic impedance, Z_{02} for the second line is 75Ω , we will have a reflection coefficient,

Equation:

$$\begin{aligned}\Gamma_{V_2} &= \frac{R_{L2} - Z_{02}}{R_{L2} + Z_{02}} \\ &= \frac{50 - 75}{50 + 75} \\ &= -0.2\end{aligned}$$

Thus a -0.75V signal is reflected off of the second load [\[link\]](#).

Exercise:

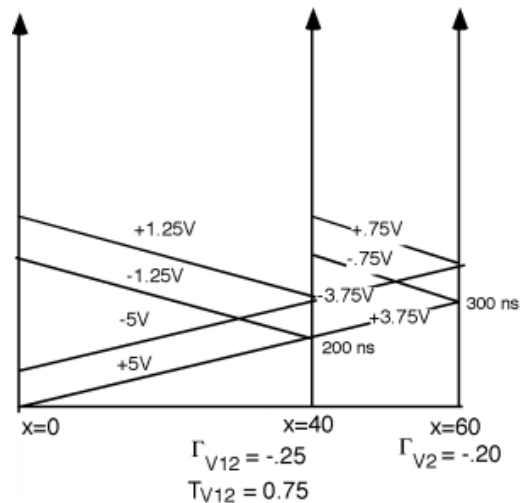
Problem:

What is the magnitude of the voltage which is developed across the second load?

Solution:

3 Volts!

Reflection of Transmitted Pulse



What happens to the 0.75V pulse when it gets to the "T"? Well there is another mismatch here, with a reflection coefficient $\Gamma_{V_{21}}$ given by

Equation:

$$\begin{aligned}\Gamma_{V_{21}} &= \frac{25 - 75}{25 + 75} \\ &= -0.5\end{aligned}$$

(The 50Ω resistor and the 50Ω transmission line look like a 25Ω termination to the 75Ω line) and a transmission coefficient

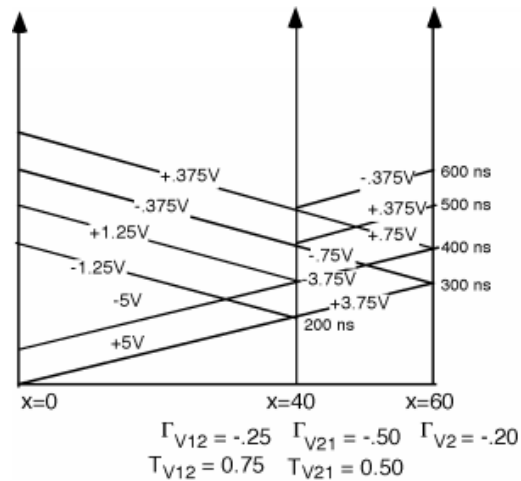
Equation:

$$T_{V_{21}} = 1 + \Gamma_{V_{21}}$$

$$= 0.5$$

and so we add to the bounce diagram [\[link\]](#).

When the Reflected Load Pulse Hits the Junction

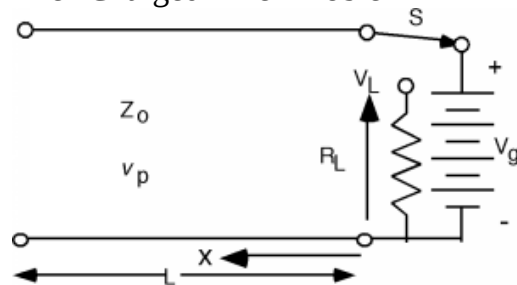


We **could** keep going, but the voltage reflected off of the second load will only be 75mV now, and so let's call it a day.

There are a couple of other interesting applications of bounce diagrams and the transient behavior of transmission lines that we might look at before we move on to other things.

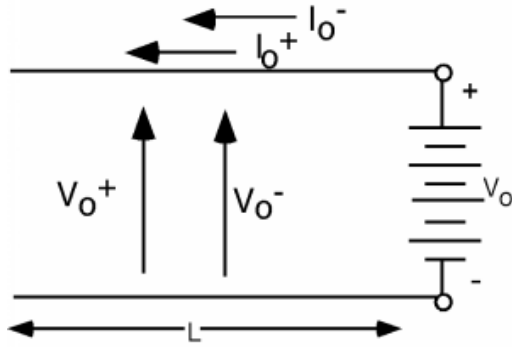
The first is called the **Charged Line Problem**. [Here](#) it is:

The "Charged Line" Problem



We have a transmission line with characteristic impedance Z_0 and phase velocity v_p . It is L long, and for some time has been connected to a battery of potential V_g [\[link\]](#). At time $t = 0$, the switch S , is thrown, which removes the battery from the circuit, and connects the line to a load resistor R_L . The question is: what does the voltage across the load resistor, V_L , look like as a function of time? This is **almost** like what we have done before, but not quite.

Initial Conditions



In the first place, we now have non-zero initial conditions. For $t < 0$ we will have both voltages and current on the line. In order to match boundary conditions, we must do more than have one voltage and one current, because the voltage on the line must be V_g , while the current flowing down the line must be 0. So, we will put in both a V^+ and a V^- and their corresponding currents. Note that x is going to the left this time. Let's forget about the switch and the load resistor for a minute and just look at the line and battery. We have two equations we must satisfy

Equation:

$$V_0^+ + V_0^- = V_g$$

and

Equation:

$$I_0^+ + I_0^- = 0$$

We can use the impedance relationship to change [\[link\]](#) to:

Equation:

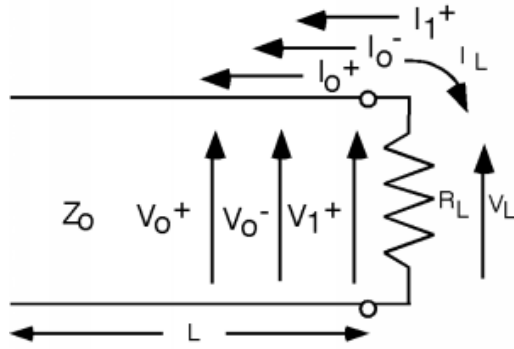
$$\frac{V_0^+}{Z_0} - \frac{V_0^-}{Z_0} = 0$$

I hope **most** of you can then see by inspection that we must have

Equation:

$$\begin{aligned} V_0^+ &= V_0^- \\ &= \frac{V_g}{2} \end{aligned}$$

OK, the switch S is thrown at $t = 0$. Now the end of the line looks like [this](#).
After the Resistor is Connected



We have anticipated the fact that we are going to need another voltage and current wave if we are going to be able to match boundary conditions when the load resistor is connected, and have added a V_1^+ and a V_1^- to the line. These are new voltage and current waves which originate at the load resistor position in order to satisfy the new boundary conditions there. Now we do KVL and KCL again.

Equation:

$$V_0^+ + V_0^- + V_1^+ = V_L$$

and

Equation:

$$\frac{V_0^+}{Z_0} - \frac{V_0^-}{Z_0} + \frac{V_1^+}{Z_0} = -\frac{V_L}{R_L}$$

We have already made the impedance substitution for the current equation in [\[link\]](#). We know what the sum and difference of V_0^+ and V_0^- are, so let's substitute in.

Equation:

$$V_g + V_1^+ = V_L$$

and

Equation:

$$\frac{V_1^+}{Z_0} = -\frac{V_L}{R_L}$$

From this we get

Equation:

$$V_L = - \left(\frac{R_L}{Z_0} V_1^+ \right)$$

which we substitute back into [\[link\]](#)

Equation:

$$V_g + V_1^+ = - \left(\frac{R_L}{Z_0} V_1^+ \right)$$

which we can solve for V_1^+

Equation:

$$\begin{aligned} V_1^+ &= - \frac{V_g}{1 + \frac{R_L}{Z_0}} \\ &= - \left(\frac{Z_0}{R_L + Z_0} V_g \right) \end{aligned}$$

The voltage on the load is given by [\[link\]](#) and is clearly just:

Equation:

$$V_L = V_g - \frac{Z_0}{R_L + Z_0} V_g$$

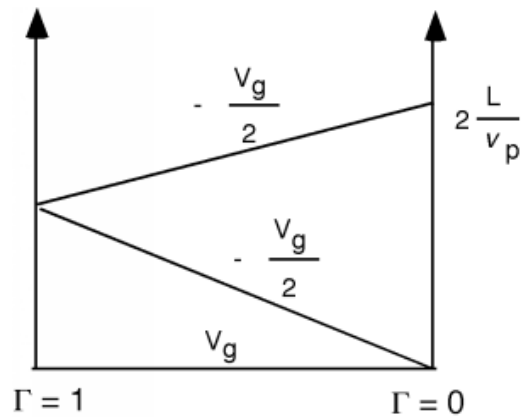
and in particular, when R_L is chosen to be Z_0 (which is usually done when this circuit is used), we have

Equation:

$$V_L = \frac{V_g}{2}$$

Now what do we do? We build a bounce diagram! Let us stay with the assumption that $R_L = Z_0$, in which case the reflection coefficient at the resistor end is 0. At the open circuit end of the transmission line Γ is +1. So we have [this](#).

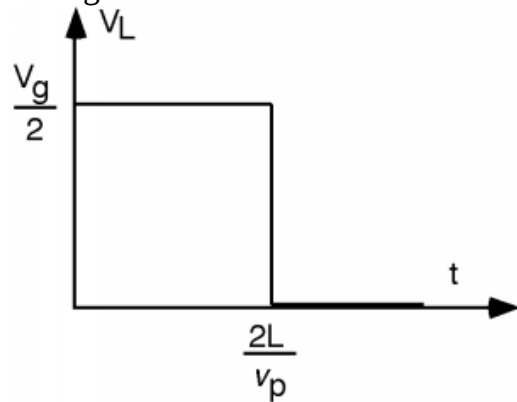
Bounce Diagram for the Charged Line Problem



Note that for **this** bounce diagram, we have added an additional voltage, V_g , on the baseline, to indicate that there is an initial voltage on the line, before the switch is thrown, and t starts on the bounce diagram.

If we concentrate on the voltage across the load, we add V_g and $-\frac{V_g}{2}$ and find that the voltage across the load resistor rises to $\frac{V_g}{2}$ at time $t = 0$ [\[link\]](#). The $-\frac{V_g}{2}$ voltage wave travels down the line, hits the open circuit, reflects back, and when it gets to the load resistor, brings the voltage across the load resistor back down to zero. We have made a pulse generator!

Voltage Across the Load Resistor



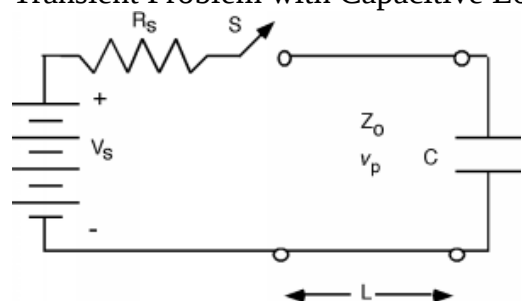
$V(t)$ across R_L

In today's digital age, this might seem like a strange way to go about creating a pulse. Imagine however, if you needed a pulse with a very large potential (100s of thousands or even millions of volts) for say, a particle accelerator. It is unlikely that a MOSFET will ever be built which is up to the task! In fact, in a field of study called **pulsed power electronics** just such circuits are used all the time. Sometimes they are built with real transmission lines, sometimes they are built from discrete inductors and capacitors,

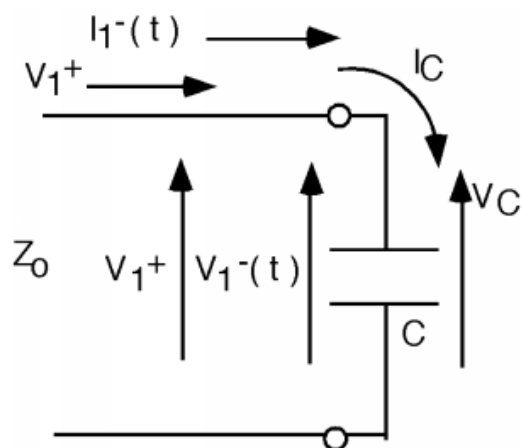
hooked together just as in the [distributed parameter model](#). Such circuits are called **pulse forming networks** or PFNs for short.

Finally, just because it affords us a good opportunity to review how we got to where we are right now, let's consider the problem of a non-resistive load on the end of a line. Suppose the line is terminated with a capacitor! For simplicity, let's let $R_s = Z_0$, so when S is closed a wave $V_1^+ = \frac{V_g}{2}$ heads down the line [\[link\]](#). Let's think about what happens when it hits the capacitor. We know we need to generate a reflected signal V_1^- , so let's go ahead and put this in the [figure](#), along with its companion current wave.

Transient Problem with Capacitive Load



Initial Pulse Hits the Load



The capacitor is initially uncharged, and we know we can not instantaneously change the voltage across a capacitor (at least without an infinite current!) and so the **initial** voltage across the capacitor should be zero, making $V_1^-(0) = -V_1^+$, if we make time $t = 0$ be when the initial wave just gets to the capacitor. So, at $t = 0$, $\Gamma_V(0) = -1$. Note that we are making Γ a function of time now, as it will change depending upon the charge state of the capacitor.

The current **into** the capacitor, I_C is just $I_1^+ + I_1^-(t)$.

Equation:

$$\begin{aligned}
 I_C(0) &= I_1^+ + I_1^-(0) \\
 &= \frac{V_g}{Z_0}
 \end{aligned}$$

since

Equation:

$$\begin{aligned}
 I_1^+ &= \frac{V_1^+}{Z_0} \\
 &= \frac{V_g}{2Z_0}
 \end{aligned}$$

and

Equation:

$$\begin{aligned}
 I_1^-(0) &= -\frac{V_1}{Z_0} \\
 &= -\frac{V_g}{2Z_0}
 \end{aligned}$$

How will the current into the capacitor $I_C(t)$ behave? We have to remember the capacitor equation:

Equation:

$$\begin{aligned}
 I_C(t) &= C \frac{dV_C(t)}{dt} \\
 &= C \left(\frac{\partial (V_1^+ + V_1^-(t))}{\partial t} \right) \\
 &= C \frac{dV_1^-(t)}{dt}
 \end{aligned}$$

since V_1^+ is a constant and hence has a zero time derivative. Well, we also know that

Equation:

$$\begin{aligned}
 I_C(t) &= I_1^+ + I_1^-(t) \\
 &= \frac{V_1^+}{Z_0} - \frac{V_1^-(t)}{Z_0}
 \end{aligned}$$

So we equate [\[link\]](#) and [\[link\]](#) and we get

Equation:

$$C \frac{d V_1^-(t)}{d t} = \frac{V_1^+}{Z_0} - \frac{V_1^-(t)}{Z_0}$$

or

Equation:

$$\frac{d V_1^-(t)}{d t} + \frac{1}{Z_0 C} V_1^-(t) = \frac{1}{C} \frac{V_1^+}{Z_0}$$

which gets us back to **another** Diff-E-Q!

The homogeneous solution is easy. We have

Equation:

$$\frac{d V_1^-(t)}{d t} + \frac{1}{Z_0 C} V_1^-(t) = 0$$

for which the solution is obviously

Equation:

$$V_{1\text{homo}}^-(t) = V_0 e^{-\frac{t}{Z_0 C}}$$

After a long time, the derivative of the homogeneous solution is zero, and so the particular solution (the constant part) is the solution to

Equation:

$$\frac{1}{Z_0 C} V_{1\text{part}}^- = \frac{1}{C} \frac{V_1^+}{Z_0}$$

or

Equation:

$$V_{1\text{part}}^- = V_1^+$$

The complete solution is the sum of the two:

Equation:

$$\begin{aligned} V_1^-(t) &= V_{1\text{homo}}^-(t) + V_{1\text{part}}^- \\ &= V_0 e^{-\frac{t}{Z_0 C}} + V_1^+ \end{aligned}$$

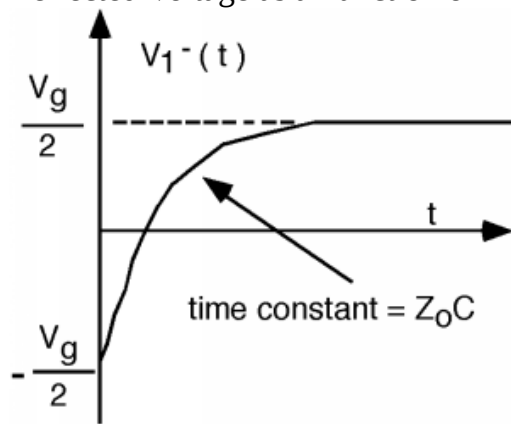
Now all we need to do is find V_0 , the initial condition. We know, however, that $V_1^-(0) = -V_1^+$, so that makes $V_0 = -2V_1^+$! So we have:

Equation:

$$\begin{aligned} V_1^-(t) &= 2V_1^+ e^{-\frac{t}{Z_0 C}} + V_1^+ \\ &= V_1^+ \left(1 - 2e^{-\frac{t}{Z_0 C}} \right) \end{aligned}$$

Since $V_1^+ = \frac{V_g}{2}$ we can plot $V_1^-(t)$ as a function of time from which we can make a [plot](#) of $\Gamma_V(t)$

Reflected Voltage as a Function of Time



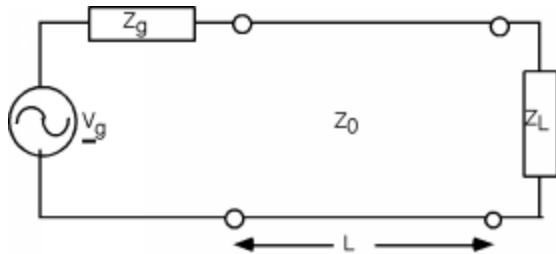
The capacitor starts off looking like a short circuit, and charges up to look like an open circuit, which makes perfect sense. Can you figure out what the shape would be of a pulse reflected off of the capacitor, given that the time constant $Z_0 C$ was short compared to the width of the pulse?

Review of Phasors

We will not always be dealing with transmission lines excited with a pulse. Although this is a good model for digital circuitry, it will not always apply. When we go to analog signals (rf, high frequency analog, etc.) we will need more tools than are available to us at this point. In the not-too-distant-past, the material we will next consider was starting to be considered passé. The rf spectrum was more or less filled up, and the watchword was "digital". Now, in the new age of wireless communication, cell phones, and rf Local Area Networks, demand for engineers who understand ac behavior on transmission lines and who can design systems which work well with rf signals are very much in demand. Pay heed to what we say here, and you might well find yourself with many lucrative job offers in the future.

To begin, we want to consider a transmission line which is being excited with an oscillating source [\[link\]](#).

Sinusoidal Excitation of a Loaded Transmission Line



The usual set-up includes a source, with a sinusoidal output, a source impedance Z_g a transmission line with impedance Z_0 , L meters long, and a load of impedance Z_L at the end.

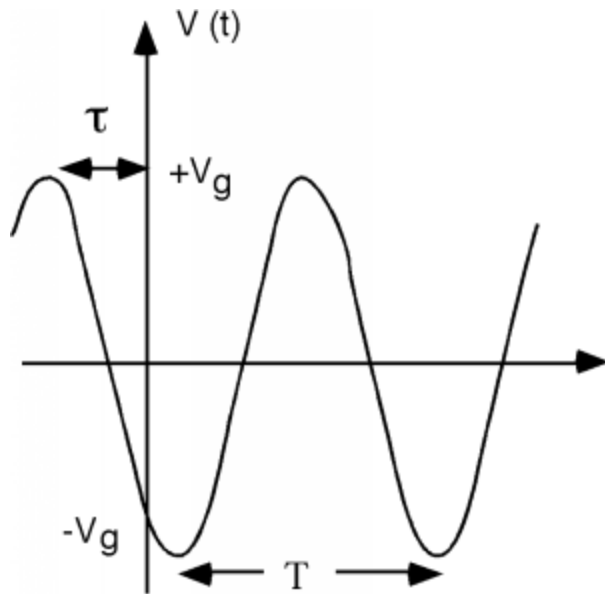
Let's look at the source first. We can describe the output waveform from the generator as

Equation:

$$V(t) = V_g \cos(\omega t + \theta)$$

Which when plotted looks like [\[link\]](#).

Excitation Waveform



The oscillating waveform has a period T and its angular frequency ω is given as

Equation:

$$\begin{aligned}\omega &= \frac{2\pi}{T} \\ &= 2\pi f\end{aligned}$$

The angle, θ , which specifies how much the wave is leading a cosine function with zero off-set is given by

Equation:

$$\theta = 2\pi \frac{\tau}{T}$$

What we **do not** want to do, is carry a bunch of sine and cosine functions around with us everywhere. Once we start multiplying and dividing, the trig turns into a big mess, and gets in the way of our understanding of what is going on. The way we deal with this, as every good 242 student knows, is to introduce **phasors**.

Since we know from **Euler's Identity**

Equation:

$$V_g e^{i(\omega t + \theta)} = V_g (\cos(\omega t + \theta) + i \sin(\omega t + \theta))$$

If we take a real part of $V_g e^{i(\omega t + \theta)}$ we will extract the voltage waveform we desire. We will re-write this function as

Equation:

$$V_g e^{i(\omega t + \theta)} = V_g e^{i\theta} e^{i\omega t}$$

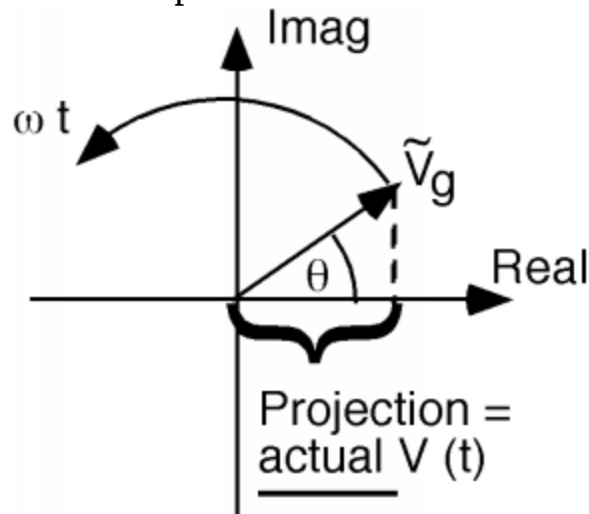
and then **define** \tilde{V}_g as the **phasor voltage** where

Equation:

$$\tilde{V}_g = V_g e^{i\theta}$$

Note that \tilde{V}_g is a complex quantity, with both a magnitude $|V_g|$ and a phase angle θ . In order to retrieve a real voltage signal from a phasor, we have to multiply the phasor by $e^{i\omega t}$ and then take the real part. Note that this is the same thing as plotting the phasor on the complex plane, and then observing the projection of the phasor on the real axis, as the phasor rotates around at a rate ωt [\[link\]](#).

Phasor Representation



This method of visualization will sometimes help make results seem a little easier to understand, or at least check for reasonableness.

A/C Line Behavior

If we are going to try to use phasors on a transmission line, then we have to allow for spatial variation as well. This is simple to do, if we just let the phasor be a function of x , so we have $\tilde{V}(x)$. How the phasor varies in x is one of the things we now have to find out.

Let's start with the **Telegrapher's Equations** again.

Equation:

$$\frac{\partial V(x, t)}{\partial x} = (-L) \frac{\partial I(x, t)}{\partial t}$$

Equation:

$$\frac{\partial I(x, t)}{\partial x} = (-C) \frac{\partial V(x, t)}{\partial t}$$

For $V(x, t)$ we can now substitute $\tilde{V}(x)e^{i\omega t}$ and for $I(x, t)$ we plug in $\tilde{I}(x)e^{i\omega t}$. So we get:

Equation:

$$\frac{\partial \left(\tilde{V}(x)e^{i\omega t} \right)}{\partial x} = (-L) \frac{\partial \left(\tilde{I}(x)e^{i\omega t} \right)}{\partial t}$$

and

Equation:

$$\frac{\partial \left(\tilde{I}(x)e^{i\omega t} \right)}{\partial x} = (-C) \frac{\partial \left(\tilde{V}(x)e^{i\omega t} \right)}{\partial t}$$

We take the derivative with respect to time, which brings down a $i\omega$ and then we cancel the $e^{i\omega t}$ from both sides of each equation:

Equation:

$$\frac{\partial \tilde{V}(x)}{\partial x} = - \left(i\omega L \tilde{I}(x) \right)$$

and

Equation:

$$\frac{\partial \tilde{I}(x)}{\partial x} = - \left(i\omega C \tilde{V}(x) \right)$$

Viola! In one simple motion, we have completely eliminated the time variable, t , from our equations! It is not really gone, of course, for once we figure out what $\tilde{V}(x)$ is, we have to multiply it by $e^{i\omega t}$ and then take the real part before we can extract once again, the actual $V(x, t)$ that we want. Nonetheless, insofar as the telegrapher's equations are concerned, t has disappeared from the radar screen.

To solve these we do just as we did with the transient problem. We take a derivative with respect to x of [\[link\]](#), which gives us a $\frac{\partial \tilde{I}(x)}{\partial x}$ on the right hand side, for which we can substitute [\[link\]](#), which leaves us with

Equation:

$$\frac{\partial^2 \tilde{V}(x)}{\partial \text{msup}} = - \left(\omega^2 L C \tilde{V}(x) \right)$$

(- times - is +, but $ii = -1$ and so we have a - in front of the ω^2). We then re-write [\[link\]](#) as

Equation:

$$\frac{\partial^2 \tilde{V}(x)}{\partial \text{msup}} + \omega^2 L C \tilde{V}(x) = 0$$

The simplest solution to this equation is

Equation:

$$\tilde{V}(x) = V_0 e^{\pm(i\omega\sqrt{LC}x)}$$

from which we can then get the actual voltage signal

Equation:

$$\begin{aligned} V(x, t) &= \tilde{V}(x) e^{i\omega t} \\ &= V_0 e^{i(\omega t \pm \omega\sqrt{LC}x)} \end{aligned}$$

Note that we could factor out an $e^{i\omega\sqrt{LC}}$, from the exponent, which, since it is just a constant, we could include in V_0 (and call it V'_0 , switch the order of x and t , and write [\[link\]](#) as

Equation:

$$V(x, t) = V'_0 e^{i\left(x \pm \frac{1}{\sqrt{LC}}t\right)}$$

which looks a lot like the "general" $f(x \pm vt)$ solution we were talking about [earlier](#)!

The number $\omega\sqrt{LC}$ is special. It is usually represented with a Greek letter β and is called the **propagation coefficient**. Thus we have

Equation:

$$V(x, t) = V_0 e^{i(\omega t \pm \beta x)}$$

As previously, a point on the wave of constant phase requires that the argument inside the parenthesis remains constant. Thus if $V(x_1, t_1)$ is going to equal $V(x_2, t_2)$ (i.e. what was at point x_1 at t_1 is now at x_2 at time t_2 it must be that

Equation:

$$\omega t_1 \pm \beta x_1 = \omega t_2 \pm \beta x_2$$

or

Equation:

$$\frac{x_2 - x_1}{t_2 - t_1} = \frac{\Delta(x)}{\Delta(t)} = \pm \left(\frac{\omega}{\beta} \right) = \pm \left(\frac{\omega}{\omega \sqrt{LC}} \right) = \pm \left(\frac{1}{\sqrt{LC}} \right) \equiv v_p$$

Which one again, defines the **phase velocity** of the wave. Other relationships to keep in mind are

Equation:

$$\beta = \frac{2\pi}{\lambda}$$

Equation:

$$\begin{aligned} \lambda &= \frac{v_p}{f} \\ &= \frac{\frac{\omega}{\beta}}{\frac{\omega}{2\pi}} \\ &= \frac{2\pi}{\beta} \end{aligned}$$

The first comes from the fact that the wave varies in x as $e^{i\beta x}$. Thus when $x = \gamma$, the wavelength, $\beta\gamma$ just increases by 2π , to get the phasor to go through one full rotation. Note also, as before, the choice of the minus sign in the \pm in [\[link\]](#) represents a wave going in the x direction, while the choice of the $+$ sign will give a wave going in the $-x$ direction. Clearly, by starting out taking the x-derivative of the equation for $I(x, t)$ we would end up with

Equation:

$$I(x, t) = I_0 e^{i(\omega t \pm \beta x)}$$

Let's consider the two phasors then, and define the voltage phasor associated with the positive going voltage wave as

Equation:

$$\tilde{V}_{\text{plus}}(x) = V^+ e^{-(i\beta x)}$$

and the negative voltage phasor as

Equation:

$$\tilde{V}_{\text{minus}}(x) = V^- e^{i\beta x}$$

We should keep in mind that both V^+ and V^- can be, and probably are, complex numbers. (From now on we will drop the little \sim over the variables because its very tedious to get it to show up with this word processor. You will just have to keep in mind that any variable we do not explicitly put inside absolute value markers (i.e. $|V^+|$) is going to be, in general, a complex number). We will, of course, have similar expressions for the positive and negative going current waves.

Let's consider the positive going current and voltage waves, and plug them into [\[link\]](#).

Equation:

$$\frac{\partial V^+ e^{-(i\beta x)}}{\partial x} = - \left(i\omega L I^+ e^{-(i\beta x)} \right)$$

The x-derivative brings down a $-(i\beta)$, the $e^{-(i\beta x)}$'s cancel, and we have

Equation:

$$V^+ = \frac{i\omega L}{i\beta} I^+$$

But, since $\beta = \omega\sqrt{LC}$ we have

Equation:

$$V^+ = \sqrt{\frac{L}{C}} I^+ \equiv Z_0 I^+$$

as we had before.

So, what has changed? Not much from the case of transients on a line. We will now assume we have a **steady state** problem. This means we turned on the generator a long time ago. We assume that it has been connected to the line long enough so that all transient behavior has died away, and that voltages and currents are not changing any more (except oscillating at frequency ω , of course).

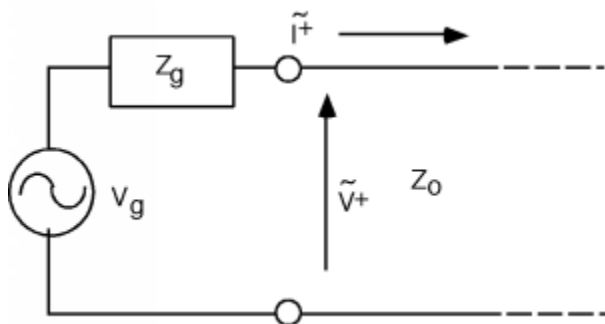
If the line is semi-infinite (or matched with a load equal to Z_0) [\[link\]](#) then it is pretty obvious that

Equation:

$$V = \frac{Z_0}{Z_0 + Z_g} V_g$$

where Z_g is the source impedance, and V_g is the source voltage phasor.

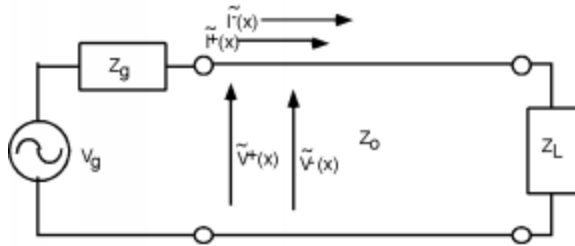
A Wave On a Semi-Infinite Line



Terminated Lines

If, on the other hand, we have a finite line, terminated with some load impedance, we have a somewhat more complicated problem to deal with [\[link\]](#).

A Finite Terminated Transmission Line



There are several things we should note **before** we head off into equation-land again. First of all, unlike the transient problems we looked at in a [previous chapter](#), there can be no more than **two** voltage and current signals on the line, just V^+ and V^- , (and I^+ and I^-). We no longer have the luxury of having V_1^+ , V_2^+ , etc., because we are talking now about a **steady state system**. All of the transient solutions which built up when the generator was first connected to the line have been summed together into just two waves.

Thus, on the line we have a single **total voltage function**, which is just the sum of the positive and negative going voltage waves

Equation:

$$V(x) = V^+ e^{-i\beta x} + V^- e^{i\beta x}$$

and a total current function

Equation:

$$I(x) = I^+ e^{-i\beta x} + I^- e^{i\beta x}$$

Note also that until we have solved for V^+ and V^- , we do not know $V(x)$ or $I(x)$ anywhere on the line. In particular, we do not know $V(0)$ and $I(0)$

which would tell us what the apparent impedance is looking into the line.

Equation:

$$\begin{aligned} Z_{\text{in}} &= Z(0) \\ &= \frac{V^+ + V^-}{I^+ + I^-} \end{aligned}$$

Until we know what kind of impedance the generator is seeing, we can not figure out how much of the generator's voltage will be coupled to the line! The input impedance looking into the line is now a function of the load impedance, the length of the line, and the phase velocity on the line. We have to solve this **before** we can figure out how the line and generator will interact.

The approach we shall have to take is the following. We will start at the **load** end of the line, and in a manner similar to the one we used previously, find a relationship between V^+ and V^- , leaving their actual magnitude and phase as something to be determined later. We can then propagate the two voltages (and currents) back down to the input, determine what the input impedance is by finding the ratio of $(V^+ + V^-)$ to $(I^+ + I^-)$, and from this, and knowledge of properties of the generator and its impedance, determine what the actual voltages and currents are.

Let's take a look at the load. We again use KVL and KCL ([\[link\]](#)) to match voltages and currents in the line and voltages and currents in the load:

Equation:

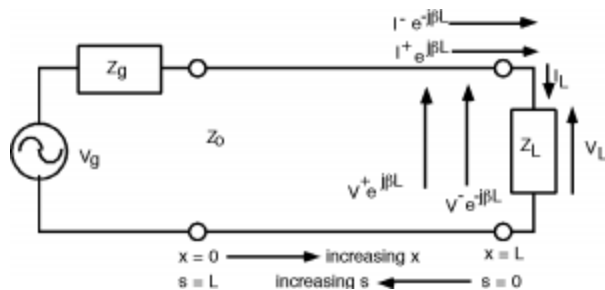
$$V^+ e^{-(i\beta L)} + V^- e^{i\beta L} = V_L$$

and

Equation:

$$I^+ e^{-(i\beta L)} + I^- e^{i\beta L} = I_L$$

Doing Kirchoff at the End of the Line



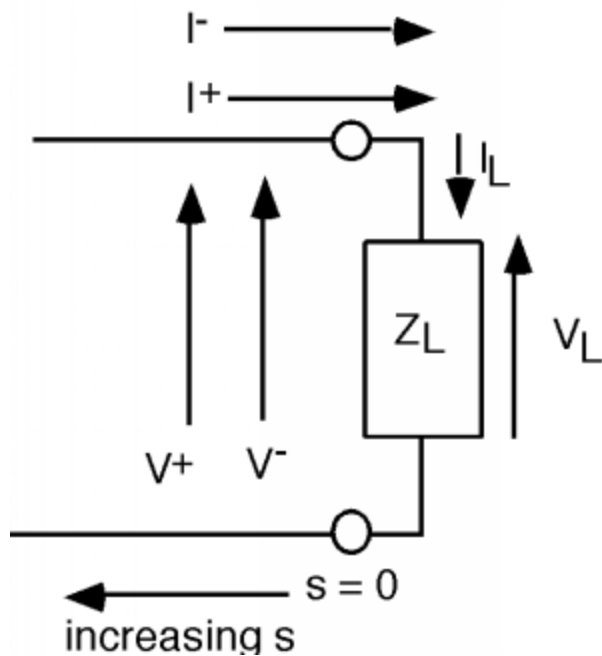
Change variables!

Now, we **could** substitute $\frac{\pm(V)}{Z_0}$ for the two currents on the line and $\frac{V_L}{Z_L}$ for I_L , and then try to solve for V^- in terms of V^+ using [\[link\]](#) and [\[link\]](#) but we can be a little clever at the outset, and make our (complex) algebra a good bit cleaner [\[link\]](#). Let's make a change of variable and let

Equation:

$$s \equiv L - x$$

$s=0$ at the Load and So the Exponentials Go Away!



This then gives us for the voltage on the line (using $x = L - s$)

Equation:

$$V(s) = V^+ e^{-(i\beta L)} e^{i\beta s} + V^- e^{i\beta L} e^{-(i\beta s)}$$

Usually, we just fold the (constant) phase terms $e^{\pm(i\beta L)}$ terms in with the V^+ and V^- and so we have:

Equation:

$$V(s) = V^+ e^{i\beta s} + V^- e^{-(i\beta s)}$$

Note that when we do this, we now have a **positive** exponential in the first term associated with V^+ and a **negative** exponential associated with the V^- term. Of course, we also get for $I(s)$:

Equation:

$$I(s) = I^+ e^{i\beta s} + I^- e^{-(i\beta s)}$$

This change now moves our origin to the **load** end of the line, and reverses the direction of positive motion. **But**, now when we plug into $e^{i\beta s}$ the value for s at the load ($s = 0$), the equations simplify to:

Equation:

$$V^+ + V^- = V_L$$

and

Equation:

$$I^+ + I^- = I_L$$

which we then re-write as

Equation:

$$\frac{V^+}{Z_0} - \frac{V^-}{Z_0} = \frac{V_L}{Z_L}$$

This is beginning to look almost exactly like a [previous chapter](#). As a reminder, we solve [\[link\]](#) for V_L

Equation:

$$V_L = \frac{Z_L}{Z_0} (V^+ - V^-)$$

and substitute for V_L in [\[link\]](#)

Equation:

$$V^+ + V^- = \frac{Z_L}{Z_0} (V^+ - V^-)$$

From which we then solve for the reflection coefficient Γ_ν , the ratio of V^- to V^+ .

Equation:

$$\frac{V^-}{V^+} \equiv \Gamma_\nu = \frac{Z_L - Z_0}{Z_L + Z_0}$$

Note that since, in general, Z_L will be complex, we can expect that Γ_ν will also be a complex number with both a magnitude $|\Gamma_\nu|$ and a phase angle θ_Γ . Also, as with the case when we were looking at transients, $|\Gamma_\nu| < 1$.

Since we now know V^- in terms of V^+ , we can now write an expression for $V(s)$ the voltage anywhere on the line.

Equation:

$$V(s) = V^+ e^{i\beta s} + V^- e^{-(i\beta s)}$$

Note again the change in signs in the two exponentials. Since our spatial variable s is going in the opposite direction from x , the V^+ phasor now goes as $i\beta s$ and the V^- phasor now goes as $-(i\beta s)$.

We now substitute in $\Gamma_\nu V^+$ for V^- in [\[link\]](#), and for reasons that will become apparent soon, factor out an $e^{i\beta s}$.

Equation:

$$\begin{aligned} V(s) &= V^+ e^{i\beta s} + \Gamma_\nu V^+ e^{-(i\beta s)} \\ &= V^+ (e^{i\beta s} + \Gamma_\nu e^{-(i\beta s)}) \\ &= V^+ e^{i\beta s} (1 + \Gamma_\nu e^{-(2i\beta s)}) \end{aligned}$$

We could have also written down an equation for $I(s)$, the current along the line. It will be a good test of your understanding of the basic equations we are developing here to show yourself that indeed

Equation:

$$I(s) = \frac{V^+ e^{i\beta s}}{Z_0} (1 - \Gamma_\nu e^{-(2i\beta s)})$$

Line Impedance

Unfortunately, since we don't know what value the phasor V^+ has, these equations do not do us a whole lot of good! One way to deal with this is to simply divide [this equation](#) into [this equation](#). That gets rid of V^+ and the $e^{i\beta s}$ and so we now come up with a **new** variable, which we shall call **line impedance**, $Z(s)$.

Equation:

$$Z(s) \equiv \frac{V(s)}{I(s)} = Z_0 \frac{1 + \Gamma_\nu e^{-2i\beta s}}{1 - \Gamma_\nu e^{-2i\beta s}}$$

$Z(s)$ represents the ratio of the total voltage to the total current anywhere on the line. Thus, if we have a line L long, terminated with a load impedance Z_L , which gives rise to a terminal reflection coefficient Γ_ν , then if we substitute Γ_ν and L into [\[link\]](#), the $Z(L)$ which we calculate will be the "apparent" impedance which we would see looking into the input terminals to the line!

There are several ways in which we can look at [\[link\]](#). One is to try to put it into a more tractable form, that we might be able to use to find $Z(s)$, given some line impedance Z_0 , a load impedance Z_L and a distance, s away from the load. We can start out by multiplying top and bottom by $e^{i\beta s}$, substituting in for Γ_ν , and then multiplying top and bottom by $Z_L + Z_0$.

Equation:

$$Z(s) = Z_0 \frac{(Z_L + Z_0)e^{i\beta s} - Z_L e^{-(i\beta s)}}{(Z_L + Z_0)e^{i\beta s} - (Z_L - Z_0)e^{-(i\beta s)}}$$

Next, we use Euler's relation, and substitute $\cos(\beta s) \pm i \sin(\beta s)$ for the exponential. Lots of things will cancel out, and if we do the math carefully, we end up with

Equation:

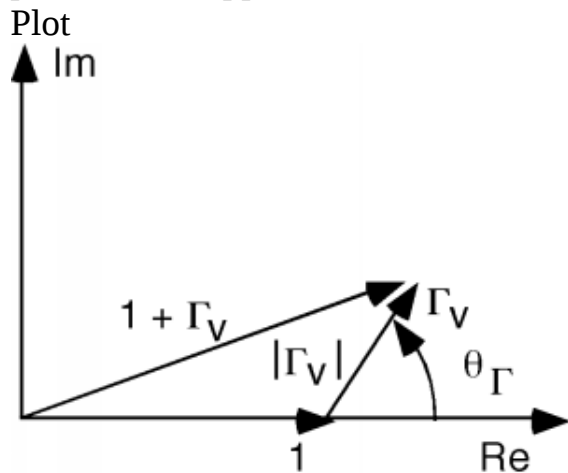
$$Z(s) = Z_0 \frac{Z_L + iZ_0 \tan(\beta s)}{Z_0 + iZ_L \tan(\beta s)}$$

For some people, this equation is more satisfying than [\[link\]](#), but for me, both are about equally opaque in terms of seeing how $Z(s)$ is going to behave with various loads, as we move down the line towards the generator. [\[link\]](#) **does** have the nice property that it is easy to calculate, and hence could be put into MATLAB or a programmable calculator. (In fact you could program [\[link\]](#) just as well for that matter.) You could specify a certain set of conditions and easily find $Z(s)$, but you would not get much insight into how a transmission line actually behaves.

Crank Diagram

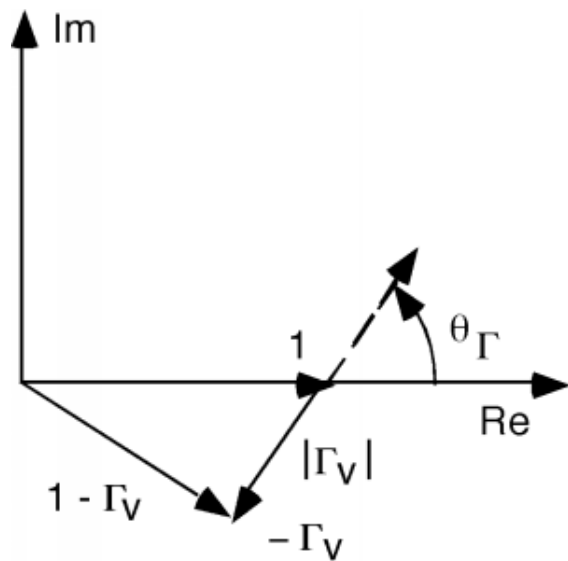
We actually still have some options open to us. One of the nicest, at least in terms of getting some insight, is called a **crank diagram**. Note that [this equation](#) is a complex equation, which requires us to take the ratio of two complex numbers; $1 + \Gamma_\nu e^{-i2\beta s}$ and $1 - \Gamma_\nu e^{-i2\beta s}$.

Let's plot these two quantities on the complex plane, starting at $s = 0$ (the load end of the line). We can represent Γ_ν , the reflection coefficient, by its magnitude and its phase, $|\Gamma_\nu|$ and φ_Γ . For the numerator we plot a 1, and then add the complex vector Γ which has a length $|\Gamma|$ and sits at an angle φ_Γ with respect to the real axis [\[link\]](#). The denominator is just the same thing, except the Γ vector points in the opposite direction [\[link\]](#).



Plotting $1 + \Gamma_\nu$

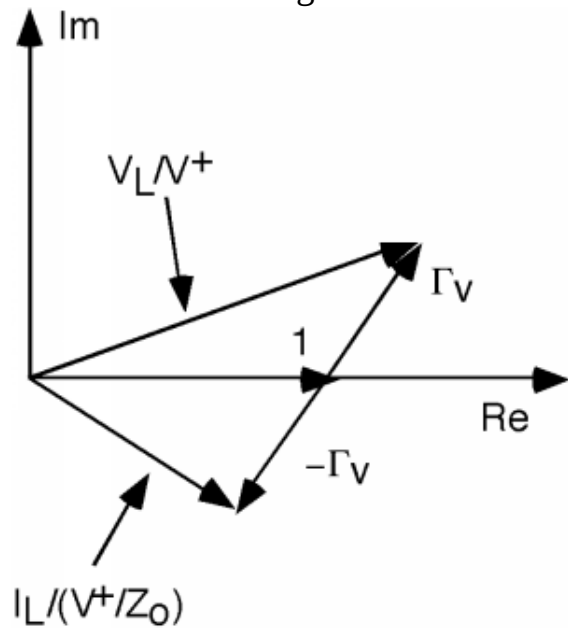
Another Plot



Plotting $1 - \Gamma_V$

The top vector is proportional to $V(s)$ and the bottom vector is proportional to $I(s)$ [\[link\]](#). Of course, for $s = 0$ we are at the load so $V(s = 0) = V_L$ and $I(s = 0) = I_L$.

Another Crank Diagram

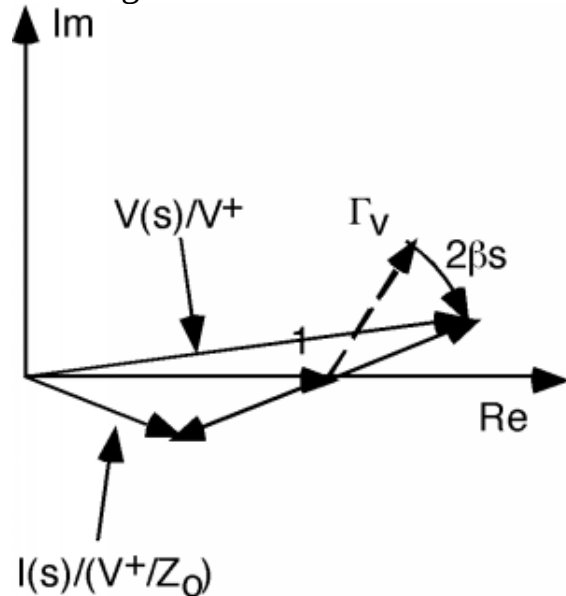


Showing that $1 + \Gamma_V = \frac{V_L}{V^+}$ and

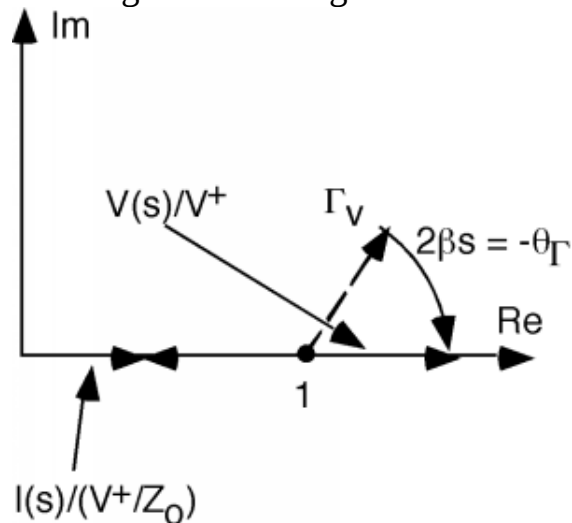
$$1 - \Gamma_v = \frac{Z_0 I_L}{V^+}$$

As we move down the line, the two " Γ " vectors rotate around at a rate of $-2\beta s$ [\[link\]](#). As they rotate, one vector gets longer and the other gets shorter, and then the opposite occurs. In any event, to get $Z(s)$ we have to divide the first vector by the second. In general, this is not easy to do, but there are **some** places where it is not too bad. One of these is when $2\beta s = -\theta_\Gamma$ [\[link\]](#).

Rotating the Phasors On the Crank Diagram



Rotating a Crank Diagram



Rotating to a V_{\max}

At this point, the voltage vector has rotated around so that it is just lying on the real axis. Obviously its length is now $1 + |\Gamma|$. By the same token, the current vector is also lying on the real axis, and has a length $1 - |\Gamma|$. Dividing one by the other, and multiplying by Z_0 gives us $Z(s)$ at this point.

Equation:

$$Z(s) = Z_0 \frac{1 + |\Gamma_\nu|}{1 - |\Gamma_\nu|}$$

Where is this point, and does it have any special meaning? For this, we need to go back to our expression for $V(s)$ in [this equation](#).

Equation:

$$\begin{aligned} V(s) &= V^+ e^{i\beta s} (1 + \Gamma_\nu e^{-2i\beta s}) \\ &= V^+ e^{i\beta s} (1 + |\Gamma_\nu| e^{i(\theta_\Gamma - 2\beta s)}) \\ &= V^+ e^{i\beta s} (1 + |\Gamma_\nu| e^{i\varphi(s)}) \end{aligned}$$

where we have substituted $|\Gamma_\nu| e^{i\theta}$ for the phasor Γ_ν and then defined a new angle $\varphi(s) = \theta_\Gamma - 2\beta s$.

Now let's find the magnitude of $V(s)$. To do this we need to square the real and imaginary parts, add them, and then take the square root.

Equation:

$$\begin{aligned} |V(s)| &= |V^+| (1 + |\Gamma_\nu| e^{i\varphi(s)}) \\ &= |V^+| \sqrt{(1 + |\Gamma_\nu| \cos(\varphi(s)))^2 + (|\Gamma_\nu|)^2 \sin^2(\varphi(s))} \end{aligned}$$

so,

Equation:

$$|V(s)| = |V^+| \sqrt{1 + 2 |\Gamma_\nu| \cos(\varphi(s)) + (|\Gamma_\nu|)^2 \cos^2(\varphi(s)) + (|\Gamma_\nu|)^2 \sin^2(\varphi(s))}$$

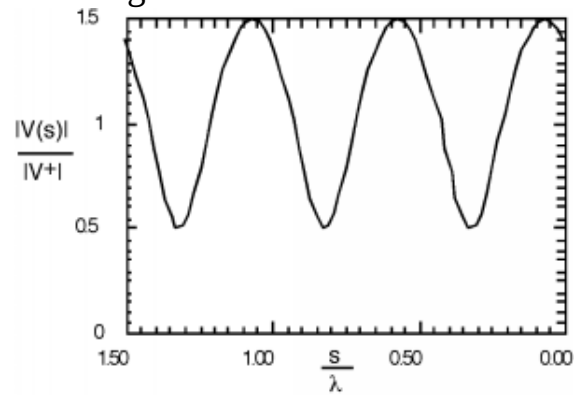
which, since $\sin^2(\cdot) + \cos^2(\cdot) = 1$

Equation:

$$|V(s)| = |V^+| \sqrt{1 + (|\Gamma_\nu|)^2 + 2|\Gamma_\nu| \cos(\varphi(s))}$$

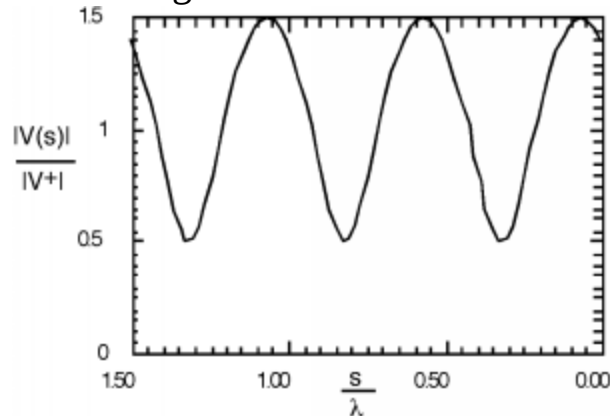
Remember, $\varphi(s)$ is an angle which changes with s . In particular, $\varphi(s) = \theta_\Gamma - 2\beta s$. Thus, as we move down the line $|V(s)|$ will oscillate as $\cos(\varphi(s))$ oscillates. A typical plot for $V(s)$ (for $|\Gamma_\nu| = 0.5$ and $\theta_\Gamma = 45^\circ$) is shown [here](#).

Standing Wave Pattern



Standing Waves/VSWR

A Standing Wave Pattern



In making [this plot](#), we have made use of the fact that the propagation constant β can also be expressed as $\frac{2\pi}{\lambda}$, and so for the independent variable, instead of showing s in meters or whatever, we normalize the distance away from the load to the wavelength of the excitation signal, and hence show distance in wavelengths. What we are showing here is called a **standing wave**. There are places along the line where the magnitude of the voltage $|V(s)|$ has a maximum value. This is where V^+ and V^- are adding up in phase with one another, and places where there is a voltage minimum, where V^+ and V^- add up out of phase. Since $|V^-| = |\Gamma_\nu| |V^+|$, the maximum value of the standing wave pattern is $1 + |\Gamma_\nu|$ times $|V^+|$ and the minimum is $1 - |\Gamma_\nu|$ times $|V^+|$. Note that anywhere on the line, the voltage is **still** oscillating at $e^{i\omega t}$, and so it is not a constant, it is just that the **magnitude** of the oscillating signal changes as we move down the line. If we were to put an oscilloscope across the line, we would see an AC signal, oscillating at a frequency ω .

A number of considerable interest is the ratio of the maximum voltage amplitude to the minimum voltage amplitude, called the **voltage standing wave ratio**, or VSWR for short. It is easy to see that:

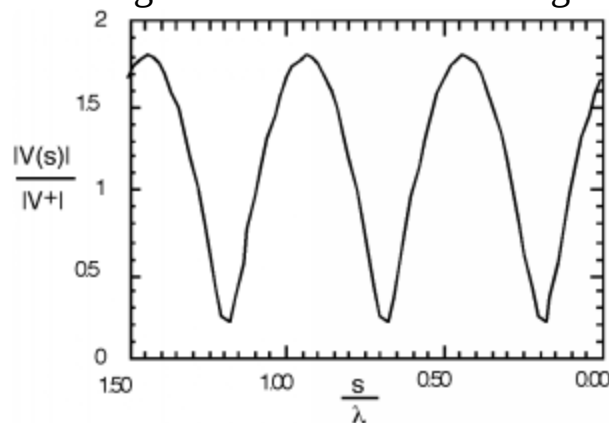
Equation:

$$\text{VSWR} = \frac{1 + |\Gamma|}{1 - |\Gamma|}$$

Note that because $|\Gamma_\nu| \in [0, 1]$, $\text{VSWR} \in [1, \infty]$.

Although [\[link\]](#) looks like the standing wave pattern is more or less sinusoidal, if we increase $|\Gamma|$ to 0.8, we see that it most definitely is not. There is also a temptation to say that the spacing between minima (or maxima) of the standing wave pattern is λ , the wavelength of the signal, but a closer inspection of either [\[link\]](#) or [\[link\]](#), shows that in fact the spacing between features is only **half** a wavelength, or $\frac{\lambda}{2}$. Why is this? Well, $\varphi(s)$ goes as $-2\beta s$ and $\beta = \frac{2\pi}{\lambda}$, and so every time s increases by $\frac{\lambda}{2}$, $\varphi(s)$ decreases by 2π and we have come one full cycle on the way $|V(s)|$ behaves.

Standing Wave Pattern with a Larger Reflection Coefficient



Now let's go back to the [Crank Diagram](#). At the position shown, we are at a voltage maximum, and $\frac{Z(s)}{Z_0}$ just equals the VSWR.

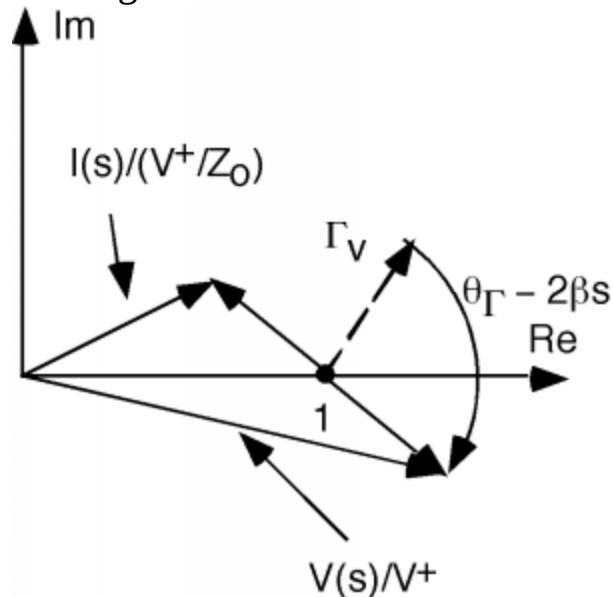
Equation:

$$\begin{aligned} \frac{Z(s_{V_{\max}})}{Z_0} &= \text{VSWR} \\ &= \frac{1+|\Gamma_\nu|}{1-|\Gamma_\nu|} \end{aligned}$$

Note also that at this particular point, that the voltage and current phasors are in phase with one another (lined up in the same direction) and hence the impedance must be **real** or resistive.

We can move further down the line, and now the $V(s)$ phasor starts shrinking, and the $I(s)$ phasor starts to get bigger [\[link\]](#).

Moving Further Down the Line



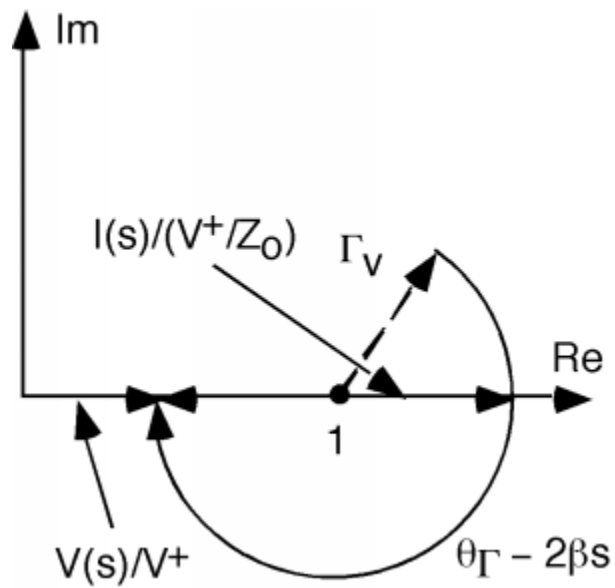
Moving further down the line
from a V_{\max}

If we move even further down the line, we get to a point where the current phasor is now at a maximum value, and the voltage phasor is at a minimum value [\[link\]](#). We are now at a voltage minimum, the impedance is again real (the voltage and current phasors are lined up with one another, so they must be in phase) and

Equation:

$$\begin{aligned} Z(s_{V_{\min}}) &= \frac{1}{\text{VSWR}} \\ &= \frac{1-|\Gamma_v|}{1+|\Gamma_v|} \end{aligned}$$

Moving Even Further Down the Line



Crank diagram at a V_{\min}

The only problem we have here is that except at a voltage minimum or maximum, finding $Z(s)$ from the crank diagram is not very straightforward, since the voltage and current are out of phase, and dividing the two vectors becomes somewhat tedious.

Bilinear Transform

There **is** a way that we can make things a good bit easier for ourselves however. The only drawback is that we have to do some complex analysis first, and look at a **bilinear transform**! Let's do one more substitution, and define another complex vector, which we can call $r(s)$:

Equation:

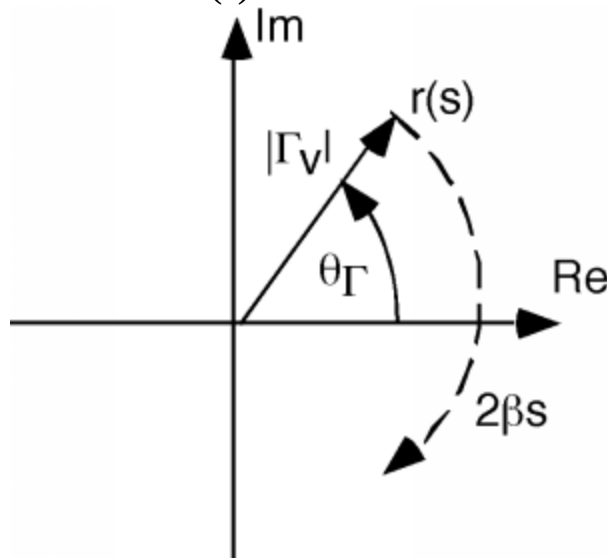
$$r(s) \equiv |\Gamma_v| e^{i(\theta_r - 2\beta s)}$$

The vector $r(s)$ is just the rotating part of the crank diagram which we have been looking at [\[link\]](#). It has a magnitude equal to that of the reflection coefficient, and it rotates around at a rate $2\beta s$ as we move down the line. For every $r(s)$ there is a corresponding $Z(s)$ which is given by:

Equation:

$$Z(s) = Z_0 \frac{1 + r(s)}{1 - r(s)}$$

The Vector $r(s)$



Now, it turns out to be easier if we talk about a **normalized impedance**, which we get by dividing $Z(s)$ by Z_0 .

Equation:

$$\frac{Z(s)}{Z_0} = \frac{1 + r(s)}{1 - r(s)}$$

which we can solve for $r(s)$

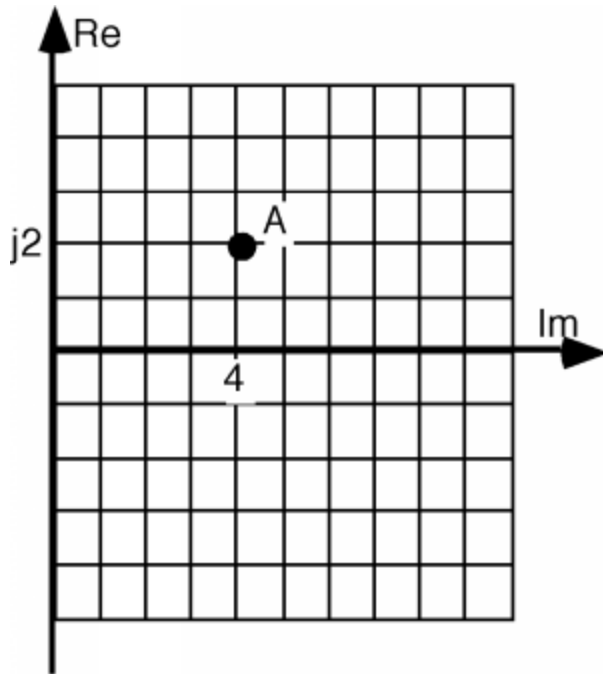
Equation:

$$r(s) = \frac{\frac{Z(s)}{Z_0} - 1}{\frac{Z(s)}{Z_0} + 1}$$

This relationship is called a **bilinear transform**. For every $r(s)$ that we can imagine, there is one and only one $\frac{Z(s)}{Z_0}$ and for every $\frac{Z(s)}{Z_0}$ there is one and only one $r(s)$. What we would like to be able to do, is find $\frac{Z(s)}{Z_0}$, given an $r(s)$. The reason for this should be readily apparent. Whereas, as we move along in s , $\frac{Z(s)}{Z_0}$ behaves in a most difficult manner (dividing one phasor by another), $r(s)$ simply rotates around on the complex plane. Given one $r(s_0)$ it is **easy** to find another $r(s)$. We just rotate around!

We shall find the required relationship in a graphical manner. Suppose I have a complex plane, representing $\frac{Z(s)}{Z_0}$. And then suppose I have some point "A" on that plane and I want to know what impedance it represents. I just read along the two axes, and find that, for the example in [\[link\]](#), "A" represents an impedance of $\frac{Z(s)}{Z_0} = 4 + 2i$. What I would like to do would be to get a grid similar to that on the $\frac{Z(s)}{Z_0}$ plane, but on the $r(s)$ plane instead. That way, if I knew one impedance (say $\frac{Z(0)}{Z_0} = \frac{Z_L}{Z_0}$ then I could find any other impedance, at any other s , by simply rotating $r(s)$ around by $2\beta s$, and then reading off the new $\frac{Z(s)}{Z_0}$ from the grid I had developed. This is what we shall attempt to do.

The Complex Impedance Plane



Let's start with [\[link\]](#) and re-write it as:

Equation:

$$\begin{aligned} r(s) &= \frac{\frac{Z(s)}{Z_0} + 1 - 2}{\frac{Z(s)}{Z_0} + 1} \\ &= 1 + \frac{-2}{\frac{Z(s)}{Z_0} + 1} \end{aligned}$$

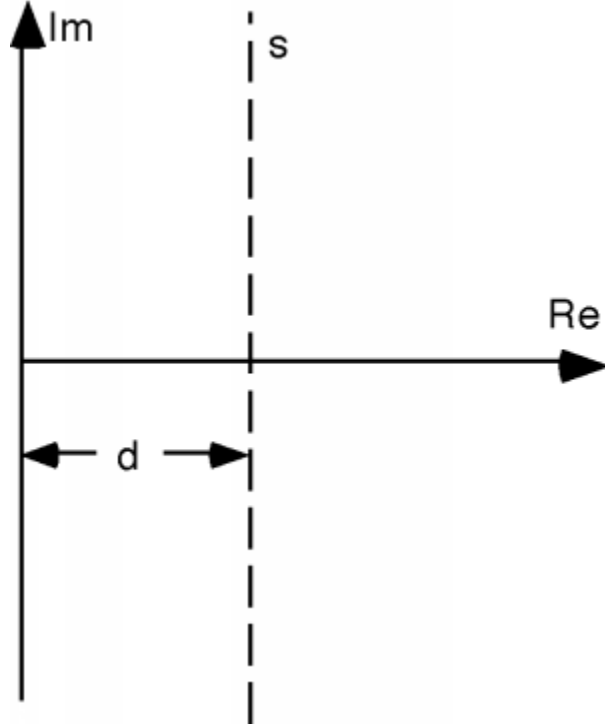
In order to use [\[link\]](#), we are going to have to interpret it in a way which might seem a little odd to you. The way we will read the equation is to say: "Take $\frac{Z(s)}{Z_0}$ and add 1 to it. Invert what you get, and multiply by -2. Then add 1 to the result." Simple isn't it? The only hard part we have in doing this is inverting $\frac{Z(s)}{Z_0} + 1$. This, it turns out, is pretty easy once we learn one very important fact.

The **one** fact about algebra on the complex plane that we need is as follows. Consider a vertical line, s , on the complex plane, located a distance d away from the imaginary axis [\[link\]](#). There are a lot of ways we could express the line s , but we will choose one which will turn out to be convenient for us. Let's let:

Equation:

$$s = d(1 - i \tan(\varphi)) \forall \varphi : \varphi \in -\frac{\pi}{2}, \frac{\pi}{2}$$

A Vertical Line, s , a Distance, d , Away From the Imaginary Axis



Now we ask ourselves the question: what is the inverse of s ?

Equation:

$$\frac{1}{s} = \frac{1}{d} \frac{1}{1 - i \tan(\varphi)}$$

We can substitute for $\tan(\varphi)$:

Equation:

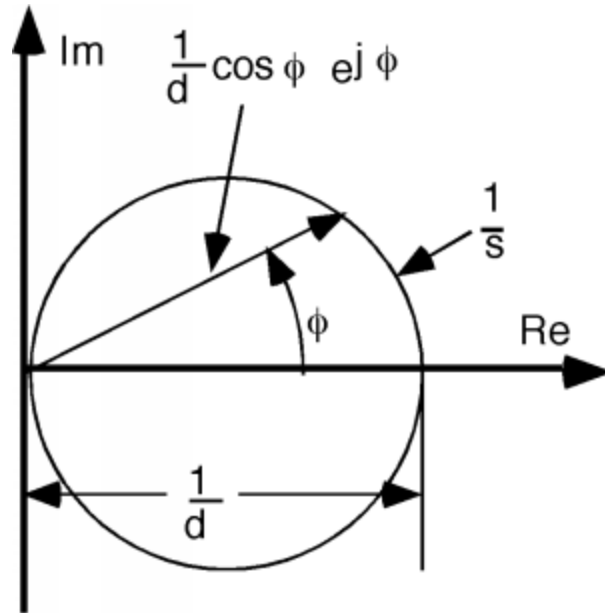
$$\begin{aligned} \frac{1}{s} &= \frac{1}{d} \frac{1}{1 - i \frac{\sin(\varphi)}{\cos(\varphi)}} \\ &= \frac{1}{d} \frac{\cos(\varphi)}{\cos(\varphi) - i \sin(\varphi)} \end{aligned}$$

And then, since $\cos(\varphi) - i \sin(\varphi) = e^{-i\varphi}$

Equation:

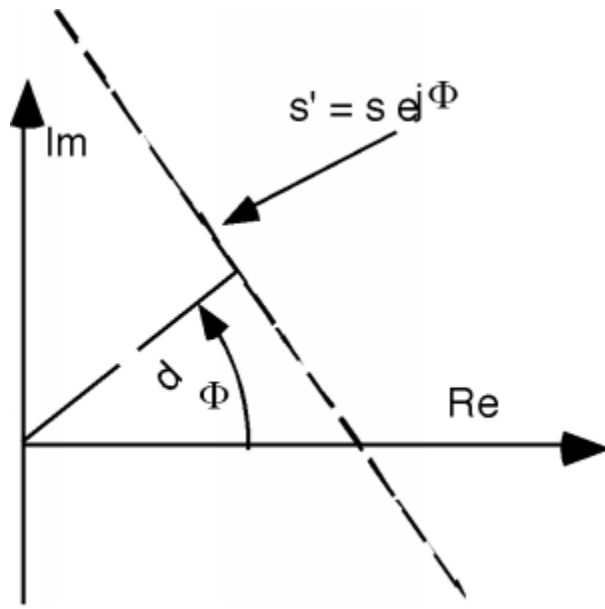
$$\begin{aligned}\frac{1}{s} &= \frac{1}{d} \frac{\cos(\varphi)}{e^{-i\varphi}} \\ &= \frac{1}{d} \cos(\varphi) e^{i\varphi}\end{aligned}$$

A Plot of $1/s$



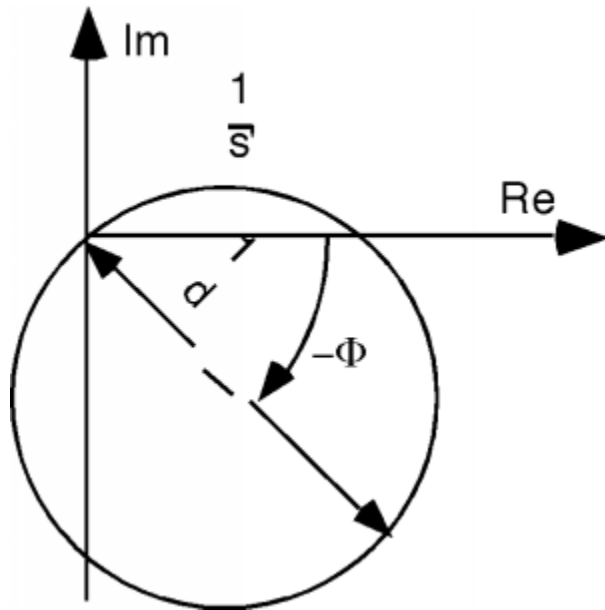
A careful look at [\[link\]](#) should allow you to convince yourself that [\[link\]](#) is an equation for a circle on the complex plane, with a diameter $\frac{1}{d}$. If s is not parallel to the imaginary axis, but rather has its perpendicular to the origin at some angle φ , to make a line s' [\[link\]](#). Since $s' = s e^{i\varphi}$, taking $\frac{1}{s}$ simply will give us a circle with a diameter of $\frac{1}{d}$, which has been rotated by an angle φ from the real axis [\[link\]](#). And so we come to the **one** fact we have to keep in mind: **"The inverse of a straight line on the complex plane is a circle, whose diameter is the inverse of the distance between the line and the origin."**

The Line s'



The line s multiplied by $e^{j\varphi}$

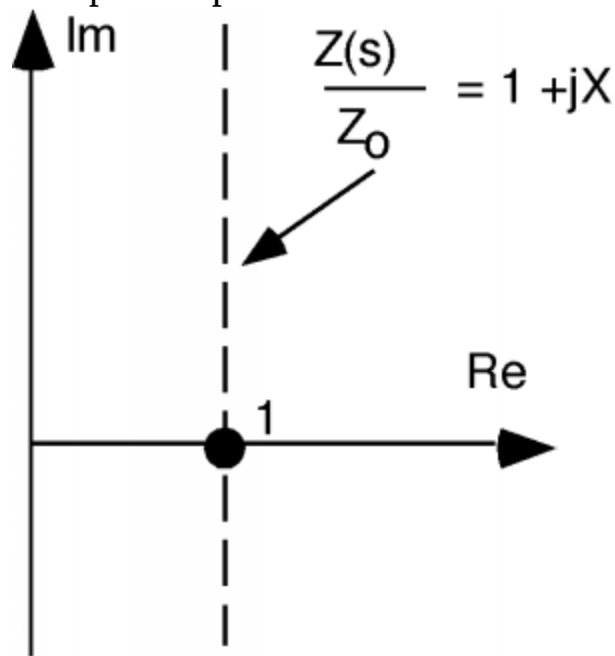
Inverse of a Rotated Line



The Smith Chart

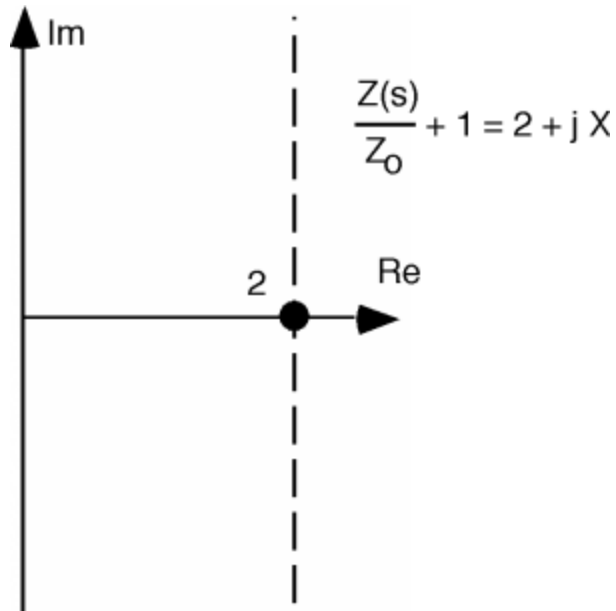
Now let's see how we can use [The Bilinear Transform](#) to get the coordinates on the $\frac{Z(s)}{Z_0}$ plane transferred over onto the $r(s)$ plane. [The Bilinear Transform](#) tells us how to take **any** $\frac{Z(s)}{Z_0}$ and generate an $r(s)$ from it. Let's start with an easy one. We will assume that $\frac{Z(s)}{Z_0} = 1 + iX$, which is a vertical line, which passes through 1, and can take on whatever imaginary part it wants [\[link\]](#).

Complex Impedence With Real Part = +1



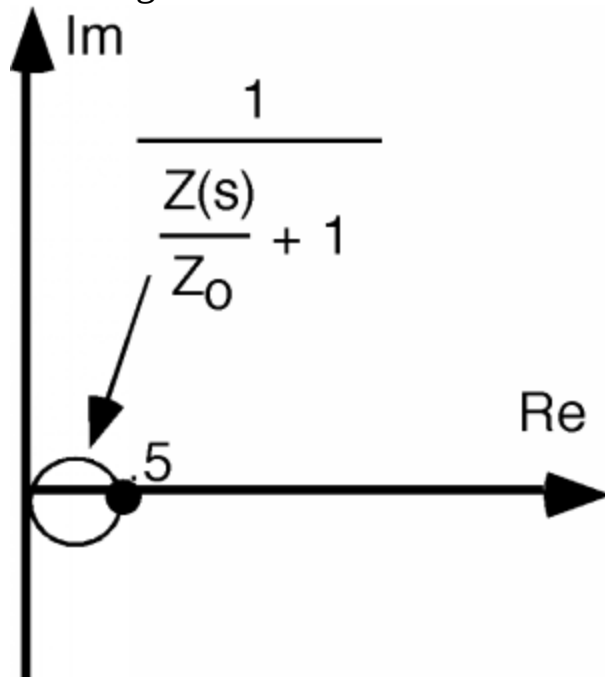
According to [The Bilinear Transform](#), the first thing we should do is add 1 to $\frac{Z(s)}{Z_0}$. This gives us the line $2 + iX$ [\[link\]](#).

Adding 1

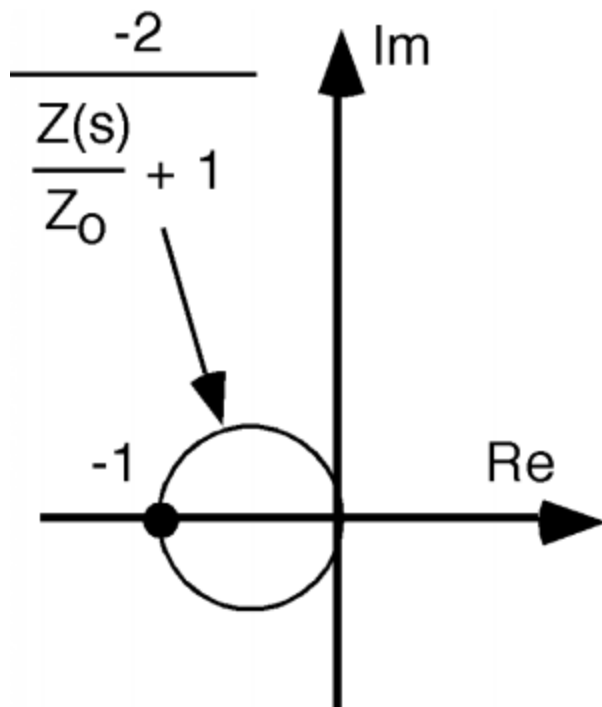


Now, we take the inverse of this, which will give us a circle, of diameter $1/2$ [\[link\]](#). Now, according to [The Bilinear Transform](#) we take this circle and multiply by -2 [\[link\]](#).

Inverting

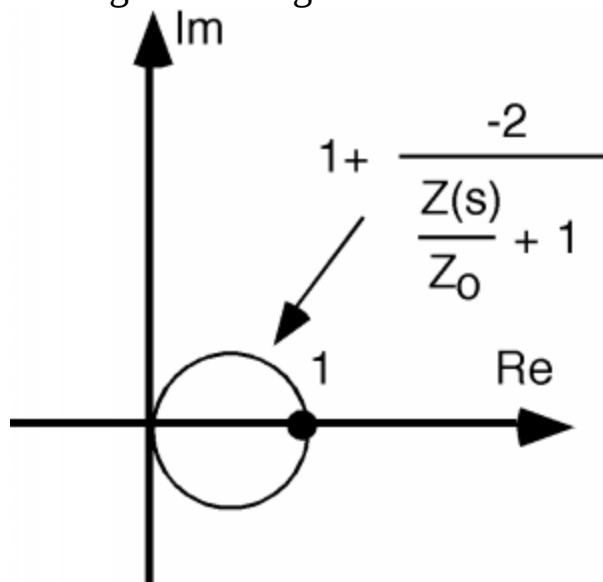


Multiplying by -2



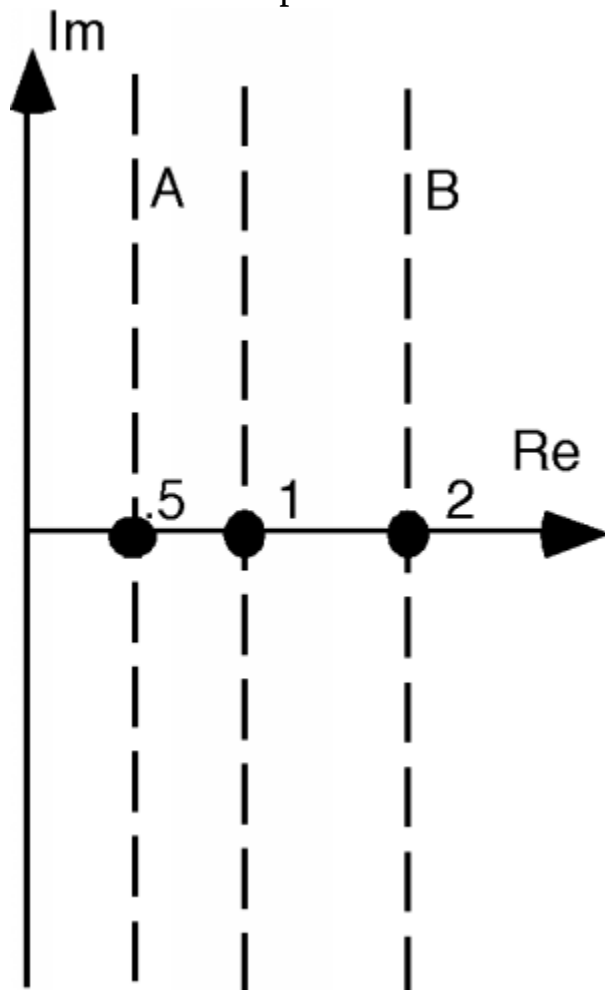
And finally, we take the circle and add +1 to it: as shown [here](#). There, we are done with the transform. The vertical line on the $\frac{Z(s)}{Z_0}$ plane that represents an impedance with a real part of +1 and an imaginary part with any value from $-(i\infty)$ to $i\infty$ has been reduced to a circle with diameter 1, passing through 0 and 1 on the complex $r(s)$ plane.

Adding 1 Once Again

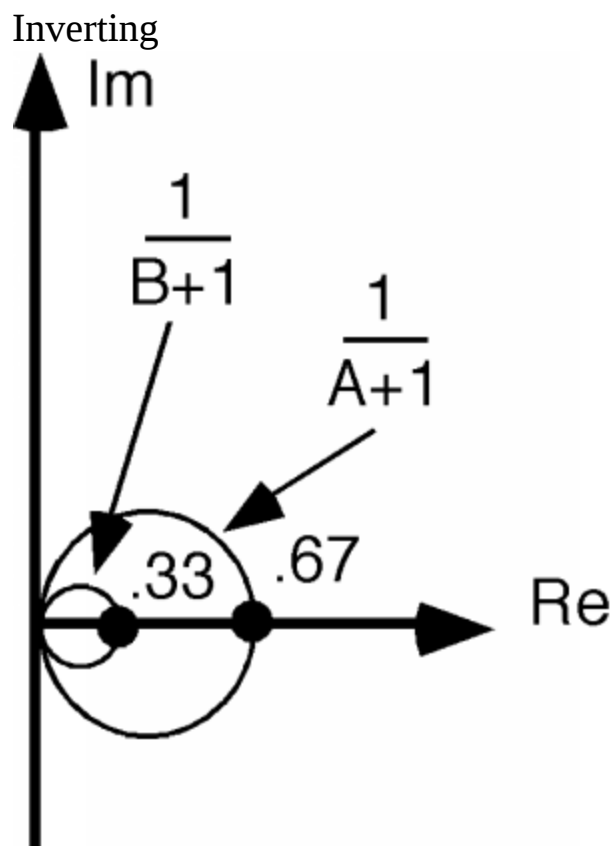
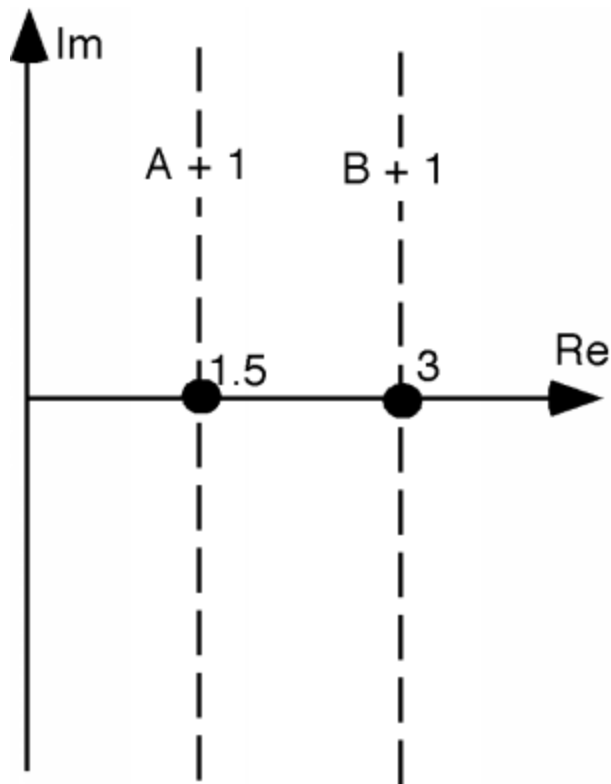


Let's do the same thing for $\frac{Z(s)}{Z_0} = 0.5 + iX$ and $\frac{Z(s)}{Z_0} = 2 + iX$. We'll call these lines A and B respectively, and just add these to the sketches we already have [\[link\]](#). Follow along with [The Bilinear Transform](#), and see if you can figure out where each of these sketches comes from. We will simply be doing the same things again: add 1 invert multiply by -2 add 1 once again. As you can see in [\[link\]](#), [\[link\]](#), [\[link\]](#), and [\[link\]](#) we get more circles. For lines inside the +1 real part, we end up with a circle that is **larger** than the +1 circle, and for lines which have a real part greater than +1, we end up with circles which are smaller in diameter than the +1 circle. All circles pass through the +1 point on the $r(s)$ plane and are tangent to one another.

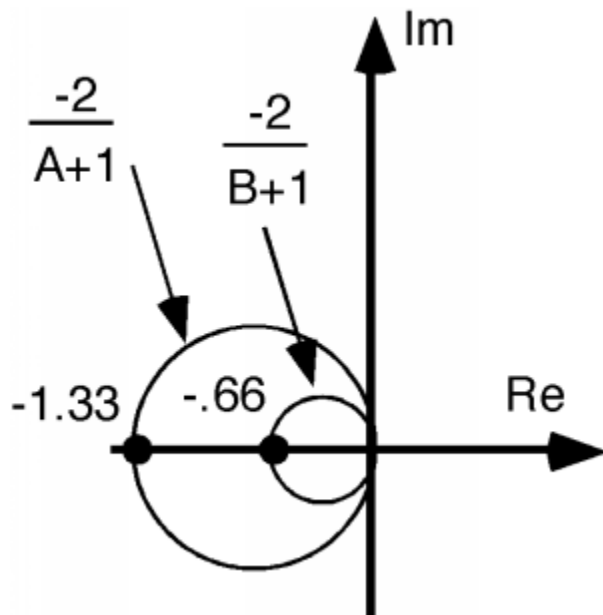
Two More Examples



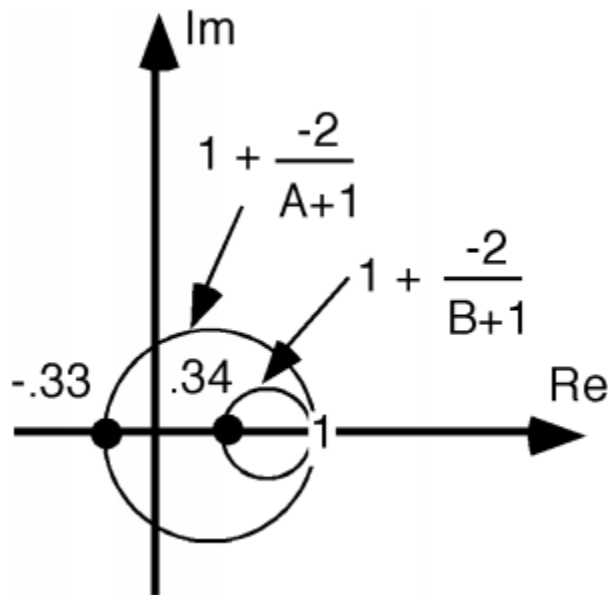
Add +1 to Each



Multiply By -2



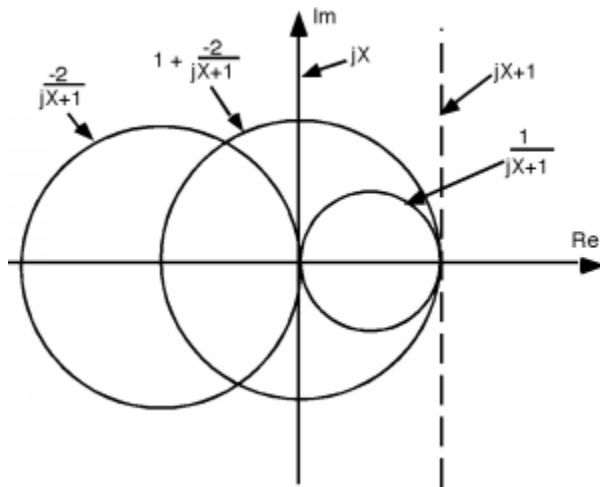
The Final Result



There are two special lines we should worry about. One is $\frac{Z(s)}{Z_0} = iX$, the imaginary axis. We will put all of the transform steps together on [\[link\]](#). We start on the axis, shift over one, get a circle with unity diameter when we invert, grow by two and flip around the imaginary axis when we multiply by -2, and then hop one to the right when +1 is added. Once again, you should work your way through the various steps to make sure you have a good understanding as to how this procedure is supposed to happen. Note

that even the imaginary axis on the $\frac{Z(s)}{Z_0}$ plane gets transformed into a circle when we go over onto the $r(s)$ plane.

Another Transform



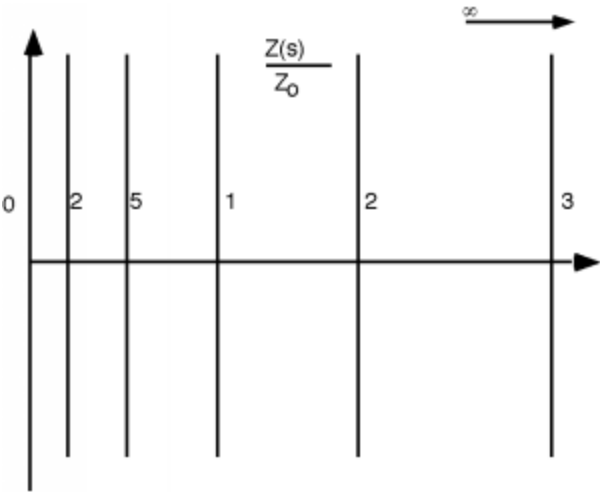
Transforming iX to the $r(s)$ plane.

The other line we should worry about is $\frac{Z(s)}{Z_0} = \infty + iX$. Now $\infty + 1 = \infty$, and $\frac{-2}{\infty} = 0.0 + 1 = 1$, and so the line $1 + iX$ gets mapped into a point at 1 when we do our transformation onto the $r(s)$ plane. Even points at ∞ on the $\frac{Z(s)}{Z_0}$ plane end up on the $r(s)$ plane, and are easily accessible!

OK, [\[link\]](#) is a plot of the $\frac{Z(s)}{Z_0}$ plane. The lines shown represent the real part of $\frac{Z(s)}{Z_0}$ that we want to transform. We run them all through [The Bilinear Transform](#), to get them onto the $r(s)$ plane. Now we have a whole family of circles, the biggest of which has a diameter of 2 (which corresponds to the imaginary axis) and the smallest of which has a diameter of 0 (which corresponds to points at ∞) [\[link\]](#). The circles all fit within one another, and since a +1 was added to every transform as the final bit of manipulation, all of the circles pass through the point +1, 0i. Circles with

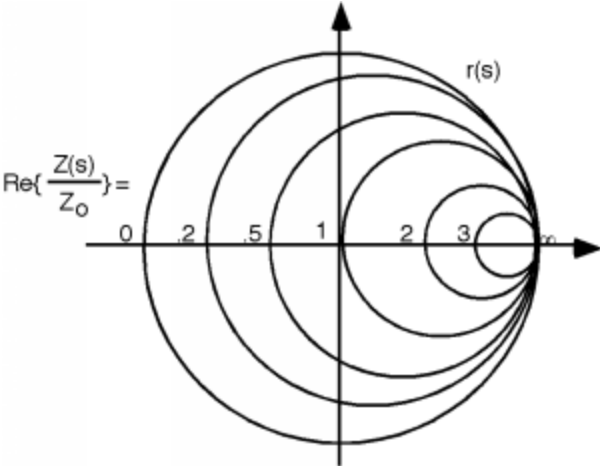
smaller diameters correspond to larger values of real $\frac{Z(s)}{Z_0}$, while the larger circles correspond to the lesser values of $\frac{Z(s)}{Z_0}$.

Other Constant Real Part Lines



Adding other constant real part
line to the $\frac{Z(s)}{Z_0}$ plane.

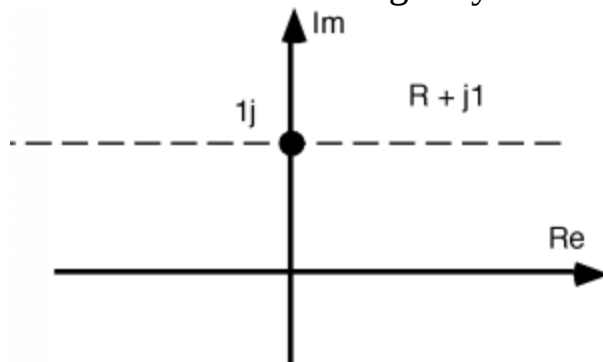
Family of Circles



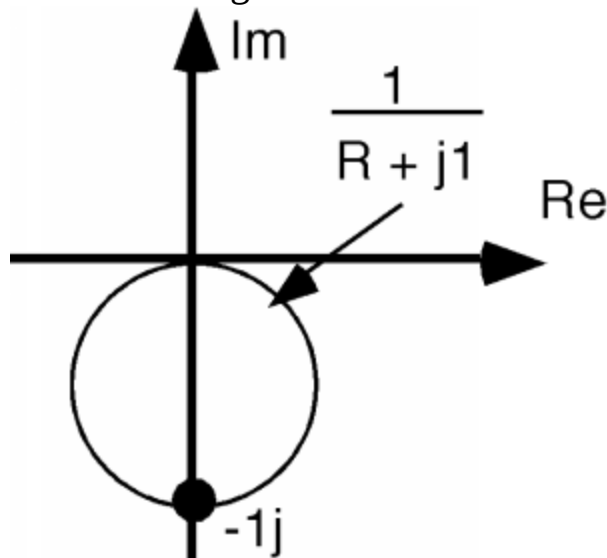
Family of $\frac{Z(s)}{Z_0}$

Well, we're half way there. Now all we have to do is find the transform for the co-ordinate lines which correspond to the imaginary part of $\frac{Z(s)}{Z_0}$. Let's look at $\frac{Z(s)}{Z_0} = R + i1$. When we add +1 to this, nothing happens! The line just slides over 1 unit, and looks just the same [\[link\]](#). Now we take its inverse. This will give us a circle, but since the line we are inverting lies at an angle of 90° with respect to the real axis, the major diameter of the circle will lie at an angle of -90° when we go through the inversion process. This gives us a circle which is lying in the $-i$ region of the complex plane [\[link\]](#).

A Line of Constant Imaginary Part

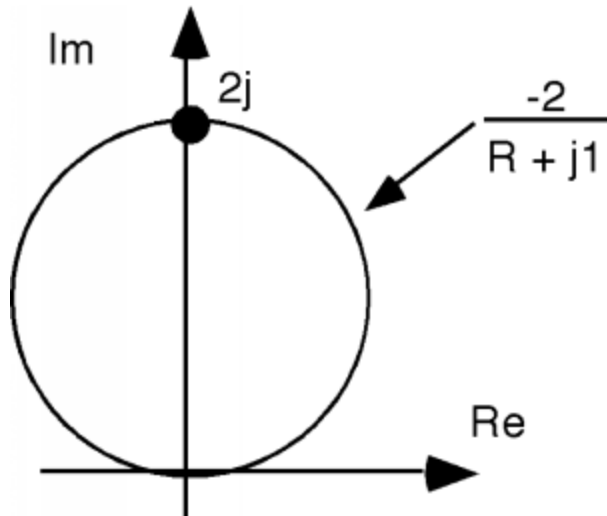


After Inverting



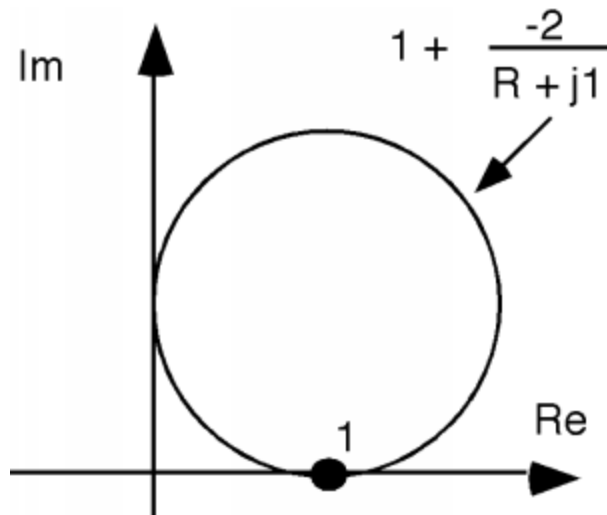
The next thing we do is to take this circle and multiply by -2. This will make the circle twice as large, but will also reflect it back up into the i region of the complex plane [\[link\]](#).

Multiply By -2



And, finally, we add 1 to it, which causes the circle to hop one over to the right [\[link\]](#).

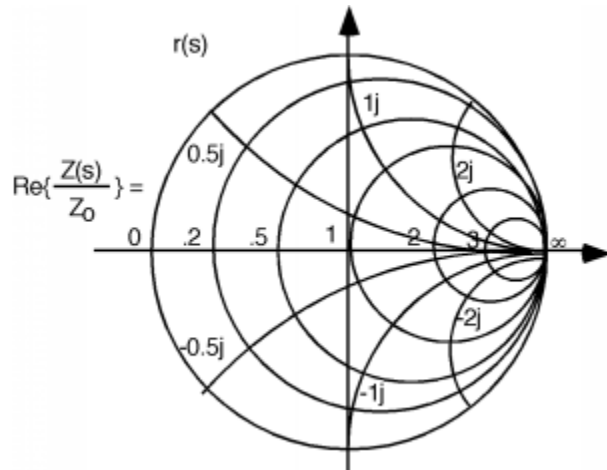
And Add 1



We can do the same thing to other lines of constant imaginary part and we can then add more circles. (Or partial circles, for it makes no sense to go beyond the $\frac{Z(s)}{Z_0} = 0$ circles, as beyond that is the region

corresponding to negative real part, which we would not expect to encounter in most transmission lines.) Take at least one of the other circles drawn [here](#) and see if you can get it to end up in about the right place.

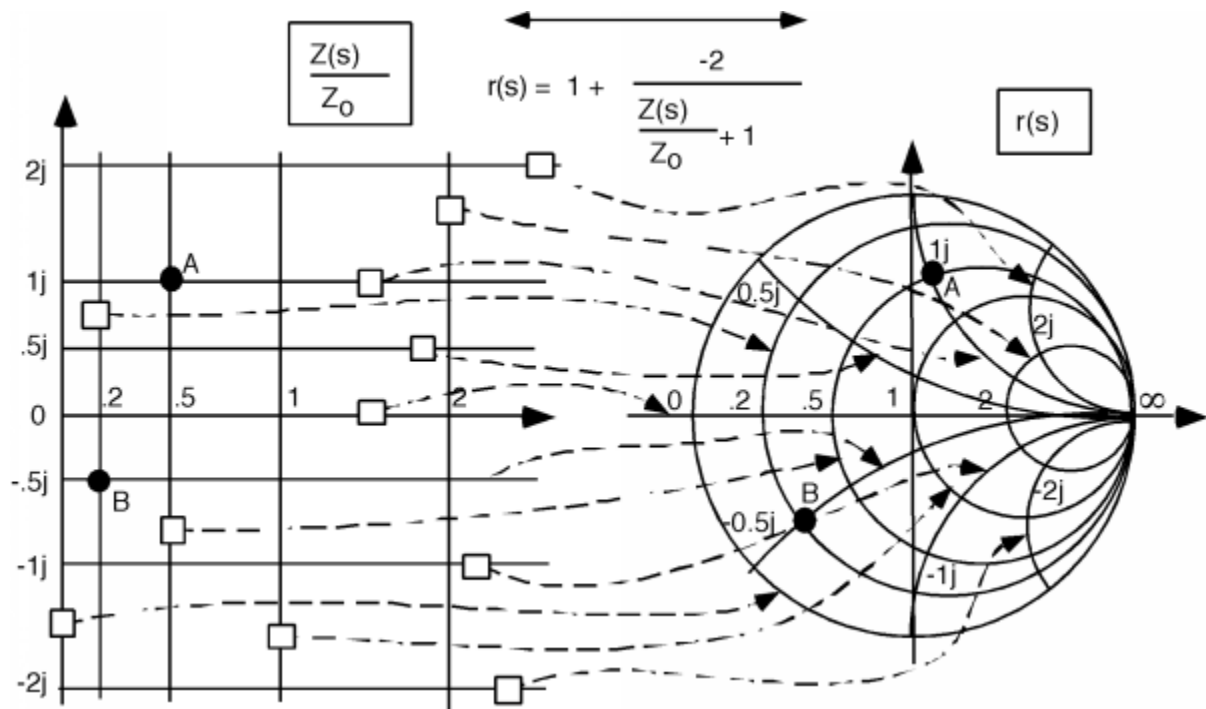
The Complete Transformation



There is one line of interest which we have to take a little care with. That is the real axis, $\frac{Z(s)}{Z_0} = 0 + iX$. This line is a distance 0 away from the origin, and so when we invert it, we get a circle with ∞ diameter. That's OK though, because that is just a straight line. So, the real axis of the $\frac{Z(s)}{Z_0}$ plane transforms into the real axis on the $r(s)$ plane.

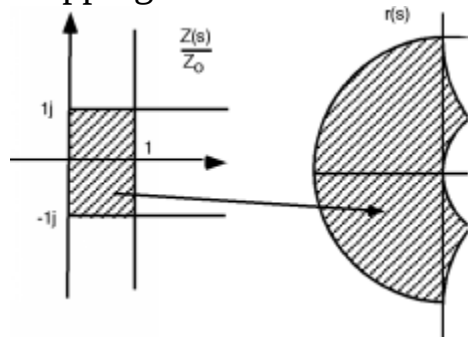
We have done a most wondrous thing! (Although you may not realize it yet.) We have taken the **entire** half plane of complex impedance $\frac{Z(s)}{Z_0}$ and mapped the whole thing into a circle with diameter 1! Let's put the two of them side by side. (Although we can't show the whole $\frac{Z(s)}{Z_0}$ plane of course.) These are shown [here](#), where we show how each line on $\frac{Z(s)}{Z_0}$ maps into a (curved) line on the $r(s)$ plane. Note also, that for every point on the $\frac{Z(s)}{Z_0}$ plane ("A" and "B") there is a corresponding point on the $r(s)$ plane. Pick a couple more points, "C" and "D" and locate them either on the $\frac{Z(s)}{Z_0}$ plane, or the $r(s)$ plane, and then find the corresponding point on the other plane.

The Mapping



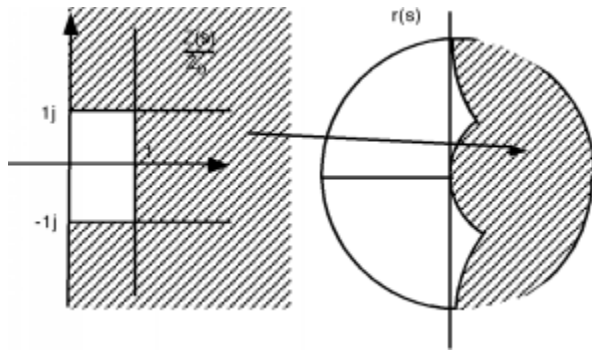
Note that the mapping is not very uniform. All of the region where either the real or imaginary part of $\frac{Z(s)}{Z_0}$ is 1 (a small square on $\frac{Z(s)}{Z_0}$ maps into a major fraction of the $r(s)$ plane [\[link\]](#) whereas all the rest of the $\frac{Z(s)}{Z_0}$ plane, all the way out to infinity in three directions (∞ , $i\infty$, and $-(i\infty)$) map into the rest of the $r(s)$ circle [\[link\]](#).

Mapping



Mapping 1, 1i

Mapping the Rest

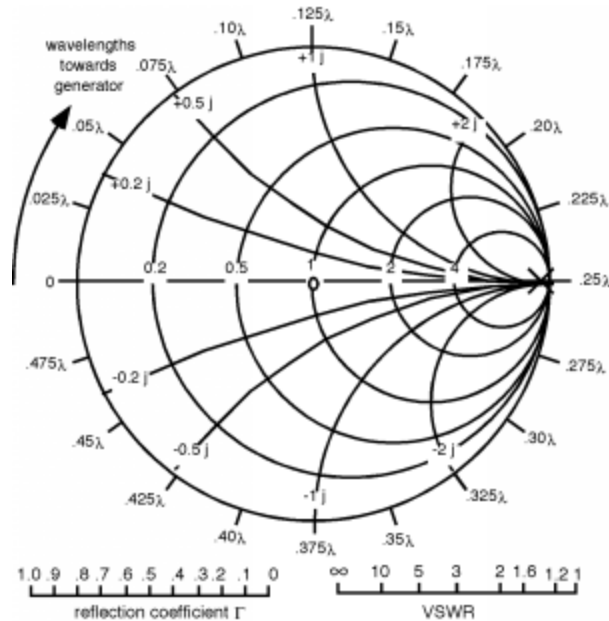


This graph or transformation is called a **Smith Chart**, after the Bell Labs worker who first thought it up. It is a most useful and powerful graphical solution to the transmission line problem. In [Introduction to Using the Smith Chart](#) we will spend a little time seeing how and why it can be so useful.

Introduction to Using the Smith Chart

Using the [Smith Chart](#), we will investigate some of the application and uses of the **Smith Chart**. For the text, we will use my new "mini Smith Chart" which is reproduced below. Clearly, there is not much detail here, and our answers will not be as accurate as they would be if we used a full size chart, but we want to get ideas across here, not the best number possible, and with the small size, we won't run out of paper before everything is done.

The Smith Chart



Note that we have a couple of "extras" on the chart. The two scales at the bottom of the chart can be used to either set or measure radial variables such as the magnitude of the reflection coefficient Γ , or the VSWR, as it turns out that in practice, what one can actually measure on a line is the VSWR. Remember, there is a direct relationship between the VSWR and the magnitude of the reflection coefficient.

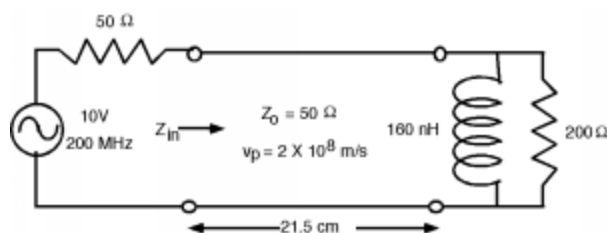
Equation:

Equation:

Since Γ , once we have the VSWR, we have Γ and so we know how big a circle we need on the Smith Chart in order to go from one place to the next. Note also that there is a scale around the outside of the chart which is given in fractions of a wavelength. Since Γ rotates around at a rate $\frac{2\pi}{\lambda}$ and Γ —, we could either show distance in cm or something, and then change the scale whenever we change wavelength. Or, we could just use a distance scale in λ , and measure all distances in units of the wavelength. This is what we shall do. Since the rate of rotation is $\frac{2\pi}{\lambda}$, one trip around the Smith Chart is the same as going one half of a wavelength down the line. Rotation in a clockwise direction is the same as moving away from the load towards the generator, while motion along the line in the other direction (towards the load) calls for counterclockwise rotation. The scale is, of course, a relative one, and so we will have to re-set our zero, depending upon where the load etc. are really located. This will all become clearer as we do an example.

Let's start out with the simplest thing we can, with just a generator, line and load [\[link\]](#). Our task will be to find the input impedance, Z_{in} , for the line, so that we can figure the input voltage.

Transmission Line Problem



For this first problem, we are going to start out with all the basics. In later examples, we probably will only give lengths in wavelengths, and impedances in terms of Z_0 , but let's do this the whole way through.

Simple Calculations with the Smith Chart

So, what do we do for Z_L ? A quick glance at a [transmission line problem](#) shows that at the load we have a resistor and an inductor in parallel. This was done on purpose, to show you one of the powerful aspects of the Smith Chart. Based on what you know from circuit theory you would calculate the load impedance by using the formula for two impedances in parallel

$Z_L = \frac{i\omega L R}{i\omega L + R}$ which will be somewhat messy to calculate.

Let's remember the formula for what the Smith Chart represents in terms of the phasor r/s .

Equation:

$$\frac{Z_L}{Z_0} = \frac{r + js}{r - js}$$

Let's invert this expression

Equation:

$$\frac{Z_L}{Z_0} = \frac{\overline{\frac{Z_L}{Z_0}}}{\frac{Y_L}{Y_0}} = \frac{r - js}{r + js}$$

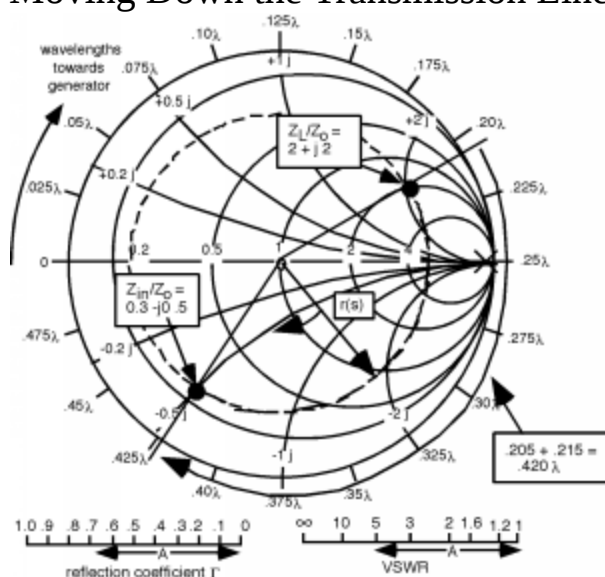
[\[link\]](#) says that if we want to get an **admittance** instead of an impedance, all we have to do is substitute r/s for r/s on the Smith Chart plane!

Equation:

$$Y = \frac{1}{Z}$$

in our case. We have two elements in parallel for the load ($Y_L = Y + iB$), so we can easily add their admittances, normalize them to Y , put them on the Smith Chart, go around (same thing as letting $r = s = r + s$) and read off $\frac{Z_L}{Z}$. For a Ω resistor, G , the conductance equals $\frac{G}{Y}$. The generator is operating at a frequency of ω , so $\omega = \pi f$ s and the inductor has a value of $i\omega L = i$ and $B = \frac{B}{i\omega L} = i$.

We plot this on the [Smith Chart](#) by first finding the real part = 0.25 circle, and then we go down onto the lower half of the chart since that is where all the negative reactive parts are, and we find the curve which represents i and where they intersect, we put a dot, and mark the location as $\frac{Y_L}{Y}$. Now to find $\frac{Z_L}{Z}$, we simply reflect half way around to the opposite side of the chart, which happens to be about $\frac{Y_L}{Y} = i$, and we mark that as well. Note that we can take the length of the line from the center of the Smith Chart to our $\frac{Z_L}{Z}$ and move it down to the Γ scale and find that the reflection coefficient has a magnitude of about 0.6. On a real Smith Chart, there is also a phase angle scale on the outside of the circle (where our distance scale is) which you can use to read off the phase angle of the reflection coefficient as well. Putting that scale on the "mini Smith Chart" would clog things up too much, but the phase angle of Γ is about 205° .
Moving Down the Transmission Line



Now the wavelength of the signal on the line is given as

Equation:

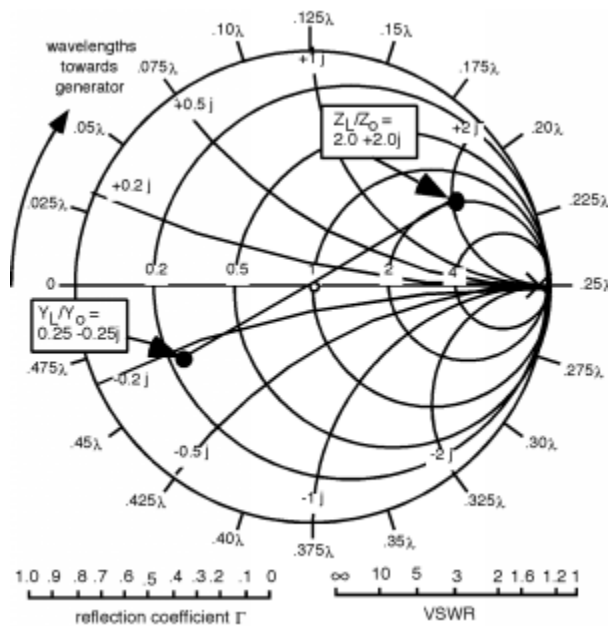
$$\lambda = \frac{\frac{v_p}{f}}{m}$$

The input to the line is located $\frac{\lambda}{2}$ or $\frac{\lambda}{4}$ away from the load. Thus, we start at $\frac{Z_L}{Z_0}$, and rotate around on a circle of constant radius a distance $\frac{\lambda}{2}$ towards the generator. To do this, we extend a line out from our $\frac{Z_L}{Z_0}$ point to the scale and read a relative distance of $\frac{\lambda}{2}$. We add $\frac{\lambda}{2}$ to this, and get $\frac{\lambda}{2}$. Thus, if we rotate around the Smith Chart, on our circle of constant radius. Since, after all, all we are doing is following r as it rotates around from the load to the input to the line. When we get to $\frac{\lambda}{2}$, we stop, draw a line out from the center, and where it intercepts the circle, we read off $\frac{Z_L}{Z_0}$ from the grid lines on the [Smith Chart](#). We find that

Equation:

$$\frac{Z}{Z_0} = i$$

Using a Smith Chart to Convert From Admittance to Impedance



Thus, $Z = 2 + j2$ ohms [\[link\]](#). Or, the impedance at the input to the line looks like a 2Ω resistor in series with a capacitor whose reactance $X_C = -2\Omega$, or, since $X_C = \frac{-j}{\omega C}$, we find that,

Equation:

$$C = \frac{1}{\omega \cdot 2}$$

To find V , there is no avoiding doing some complex math:

Equation:

$$V = \frac{1}{1 + j2}$$

Which, we write in polar notation, divide, figure the voltage and then return to rectangular notation.

Equation:

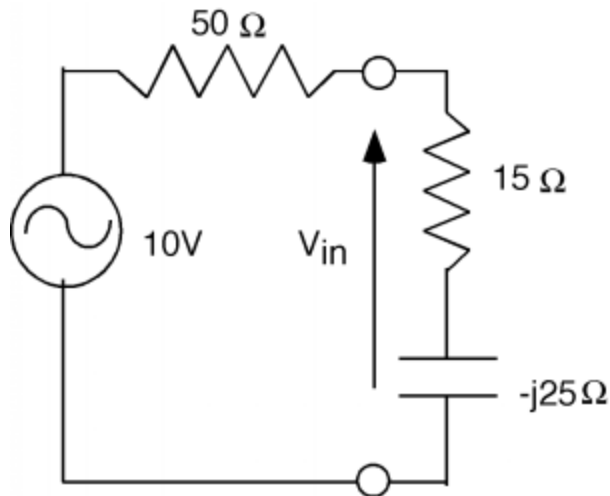
$$V = \frac{1}{\sqrt{5}}$$

Equation:

$$V$$

$$i$$

Find V_{in}



If at this point we needed to find the actual voltage phasor V we would have to use the equation

Equation:

$$V = V e^{i\beta L} \quad \Gamma V e^{-i\beta L}$$

$$V e^{i\beta L} = \Gamma V e^{i\theta_r - \beta L}$$

Where $\beta = \frac{\pi}{\lambda}$ is the propagation constant for the line as mentioned in the [last chapter](#), and L is the length of the line.

For this example, $\beta L = \frac{\pi}{\lambda} \lambda$ and $\theta_r = \Gamma$. Thus we have:

Equation:

$$V = V e^{i} \quad V e^{i}$$

Which then gives us:

Equation:

$$V = \frac{V}{e^i e^i}$$

When you expand the exponentials, add and combine in rectangular coordinates, change to polar, and divide, you will get a phasor value for V . If you do it correctly, you will find that V

Many times we don't care about V itself, but are more interested in how much power is being delivered to the load. Note that power delivered to the input of the line is also the amount of power which is delivered to the load! Finding I is easy, it's just $\frac{V}{Z}$. All we have to do is change Z to polar form.

Equation:

$$Z = |Z| e^{i\theta}$$

Equation:

$$I = \frac{V}{Z}$$

Power

You might be tempted to now say that $P_{\text{in}} = V_{\text{in}}I_{\text{in}}$, but that is incorrect for sinusoidal excitation. V_{in} and I_{in} are **phasors**! So let's digress for a second to see (or review, I hope) how to find power when the voltage and current are phasor quantities. What really matters is not the absolute phase angle of the two quantities, but rather the phase angle between them. Suppose we have a voltage phasor, V which has zero phase angle and a complex impedance $Z = |Z|e^{i\theta_z}$. Obviously, the current is given by

Equation:

$$\begin{aligned}\tilde{I} &= \frac{\tilde{V}}{\tilde{Z}} \\ &= \frac{|V|}{|Z|} e^{-\theta_z}\end{aligned}$$

To find power, we can not work just with phasors, we have to go back to the complete function of time as well so we write:

Equation:

$$V(t) = |V| \cos(\omega t)$$

Equation:

$$I(t) = \frac{|V|}{|Z|} \cos(\omega t - \theta_z)$$

Equation:

$$I(t) = |I| \cos(\omega t - \theta_z)$$

The power as a function of time is given as

Equation:

$$\begin{aligned}P(t) &= I(t)V(t) \\ &= |V| |I| \cos(\omega t) \cos(\omega t - \theta_z)\end{aligned}$$

We remember a useful trig identity:

Equation:

$$\cos(A - B) = \cos(A) \cos(B) + \sin(A) \sin(B)$$

Hence:

Equation:

$$\cos(\omega t - \theta_z) = \cos(\omega t) \cos(\theta_z) + \sin(\omega t) \sin(\theta_z)$$

which makes $P(t)$

Equation:

$$P(t) = \cos^2(\omega t) \cos(\theta_z) + \cos(\omega t) \sin(\omega t) \sin(\theta_z)$$

We are really interested in finding **average power** since energy which flows into and then back out of the line does no work for us. Clearly the second term in [\[link\]](#) (going as $\cos(\omega t) \sin(\omega t)$) has an average value of zero, and so we can forget about it. Time for one more trig identity:

Equation:

$$\cos^2(A) = \frac{1}{2} + \frac{1}{2} \cos(2A)$$

$\cos(2\omega t)$ has zero average value as well, so we are left with the following

for the average value of the power $P(t)$

Equation:

$$\begin{aligned} P(t) &= \frac{|V||I|}{2} \cos(\theta_z) \\ &= \frac{(|V|)^2}{2|Z|} \cos(\theta_z) \end{aligned}$$

Note that one useful way that people sometimes use to express this is to say

Equation:

$$P(t) = \frac{1}{2} (\tilde{V} \tilde{V}^*)$$

Back to our example: $V_{in} = 4.18 \angle 38$ and $I_{in} = 0.144 \angle 21$ Thus

Equation:

$$\begin{aligned} P_{in}(t) &= \frac{1}{2} (4.18 \times 0.144) \cos(59) \text{ Watts} \\ &= 0.155 \end{aligned}$$

As an alternative way of calculating the power into the line note that we **know** the magnitude of the current through both the capacitor and the resistor of the apparent Z_{in} . They are just two elements in series, and so they both have the same current flowing through them, namely, I_{in} . No power is dissipated in the capacitor, so we could just as well have said

Equation:

$$\begin{aligned} P_{in}(t) &= \frac{1}{2} (|I|)^2 R \\ &= \left(\frac{1}{2} 0.144^2 \right) 15 \\ &= 0.155 \end{aligned}$$

and gotten the answer in an even easier fashion! (Note that we still have to keep the factor of "1/2" to account for the time average of a sinusoidal product.) For reasons I do not understand, students have always had an aversion to finding power. It is not that hard, and in the end, is usually the "bottom line" with regard to how a system will perform. Go back over this section until it makes sense, as you may see power crop up someplace else one of these days!

Finding the Load Impedance

Let's move on to some other Smith Chart applications. Suppose, somehow, we can obtain a plot of $V(s)$ on a line with some unknown load on it. The data might look like [\[link\]](#). What can we tell from this plot? Well, $V(\max) = 1.7$ and $V(\min) = 0.3$ which means

Equation:

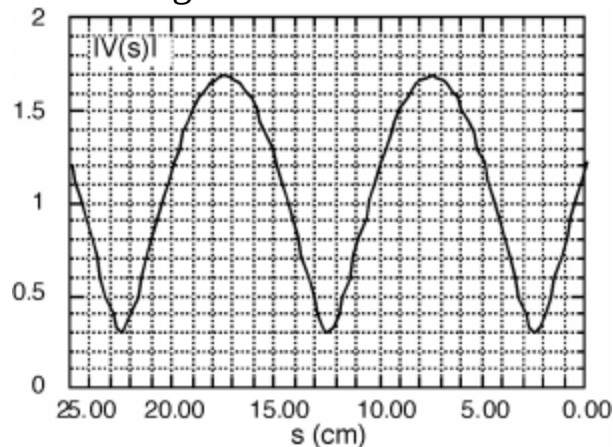
$$\begin{aligned}\text{VSWR} &= \frac{1.7}{0.3} \\ &= 5.667\end{aligned}$$

and hence

Equation:

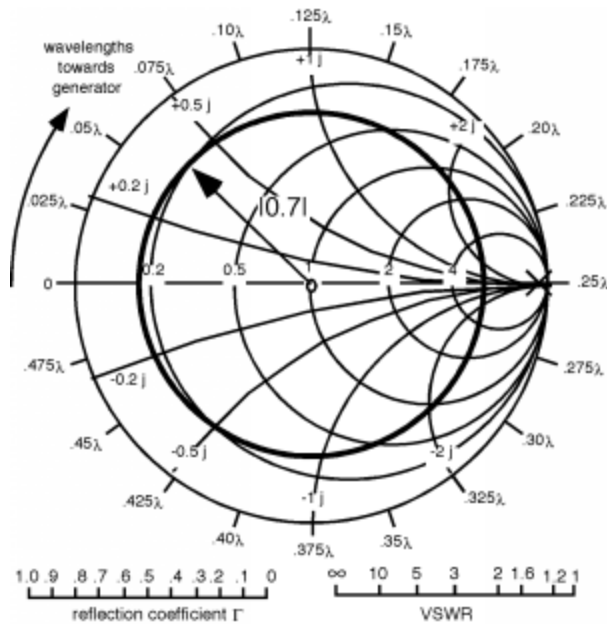
$$\begin{aligned}|\Gamma| &= \frac{\text{VSWR}-1}{\text{VSWR}+1} \\ &= \frac{4.667}{6.667} \\ &= 0.7\end{aligned}$$

A Standing Wave Pattern



Since $|r(s)| = |\Gamma|$, we can plot $r(s)$ on the Smith Chart, as shown [here](#). We do this by setting the compass at a radius of 0.7 and drawing a circle! Now, $\frac{Z_L}{Z_0}$ is **somewhere** on this circle. We just do not know where yet! There is more information to be gleaned from the VSWR plot however.

The VSWR Circle

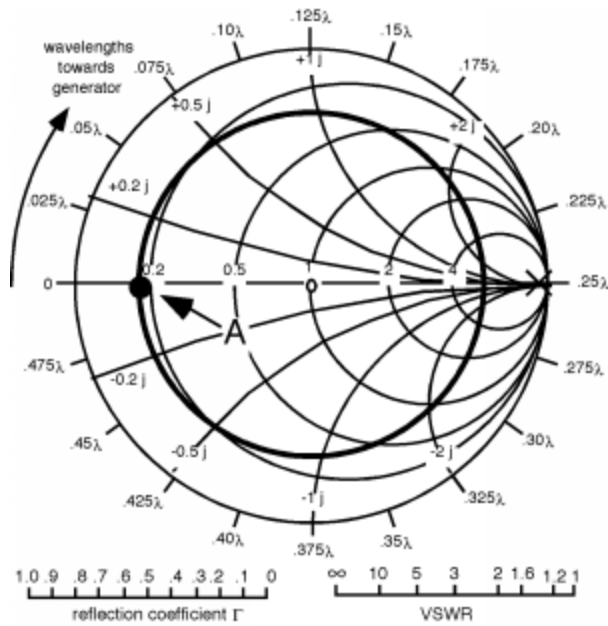


Firstly, we note that the plot has a periodicity of about 10 cm. This means that λ the wavelength of the signal on the line is 20 cm. Why? According to [this](#) equation, $|V(s)|$ goes as $\cos(\varphi(s))$ and $\varphi(s) = \theta_r - 2\beta s$ and $\beta = \frac{2\pi}{\lambda}$, thus $|V(s)|$ goes as $\cos \frac{4\pi s}{\lambda}$. Thus each $\frac{\lambda}{2}$, we are back to where we started.

Secondly, we note that there is a voltage minima at about 2.5 cm away from the load. Where on [\[link\]](#) would we expect to find a voltage minima? It would be where $r(s)$ has a phase angle of 180° or point "A" shown in [here](#). The voltage minima is **always** where the VSWR circle passes through the real axis on the left hand side. (Conversely a voltage maxima is where the circle goes through the real axis on the right hand side.) We don't really care about $\frac{Z(s)}{Z_0}$ at a voltage minima, what we want is $\frac{Z(s=0)}{Z_0}$, the normalized load impedance. This should be easy! If we start at "A" and go $\frac{2.5}{20} = 0.125\lambda$ **towards the load** we should end up at the point corresponding to $\frac{Z_L}{Z_0}$. The arrow on the mini-Smith Chart says

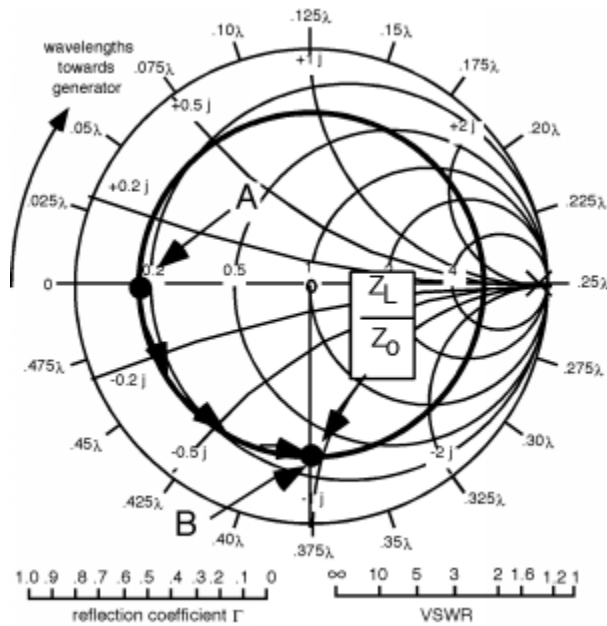
"Wavelengths towards generator" If we start at A, and want to go towards the **load**, we had better go around the opposite direction from the arrow. (Actually, as you can see on a **real** Smith Chart, there are arrows pointing in both directions, and they are appropriately marked for your convenience.)

Location of a Vmin



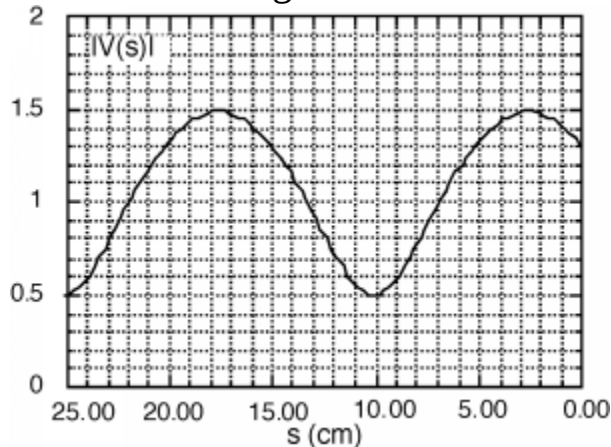
So we start at "A" go 0.125λ in a counter-clockwise direction, and mark a new point "B" which represents our $\frac{Z_L}{Z_0}$ which appears to be about $0.35 + -0.95j$ or so [\[link\]](#). Thus, the load in this case (assuming a 50Ω line impedance) is a resistor, again by co-incidence of about 50Ω , in series with a capacitor with a negative reactance of about 47.5Ω . Note that we could have started at the minima at 12.5 cm or even 22.5 cm, and then have rotated $\frac{12.5}{20} = 0.625\lambda$ or $\frac{22.5}{20} = 1.125\lambda$ towards the load. Since $\frac{\lambda}{2} = 0.5\lambda$ means one complete rotation around the Smith Chart, we would have ended up at the same spot, with the same $\frac{Z_L}{Z_0}$ that we already have! We could also have started at a maxima, at say 7.5 cm, marked our starting point on the right hand side of the Smith chart, and then we would go 0.375λ counterclockwise and again, we'd end up at "B".

Moving from Vmin to the Load

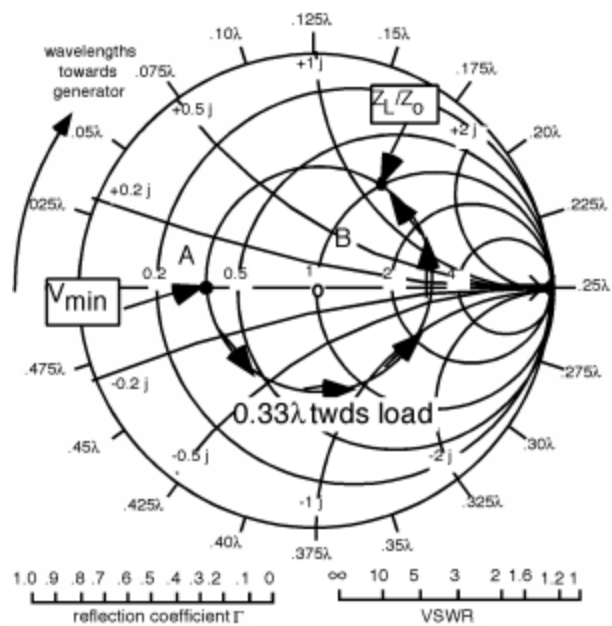


Now, [here](#) is another example. In this case the $VSWR = \frac{1.5}{0.5} = 3$, which means $|\Gamma| = 0.5$ and we get a circle as shown in [\[link\]](#). The wavelength $\lambda = 2 \times (25 - 10) = 30$ cm. The first minima is thus a distance of $\frac{10}{30} = 0.333\lambda$ from the load. So we again start at the minima, "A" and now rotate as distance 0.333λ **towards the load**.

Another Standing Wave Pattern



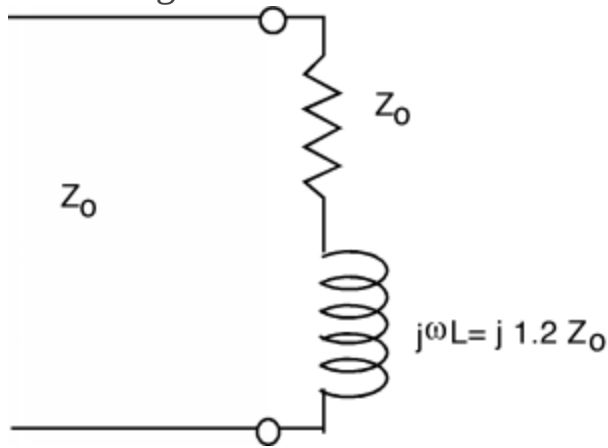
The VSWR Circle



Matching

This gets us to "B", and we find that $\frac{Z}{Z_0}$ interesting

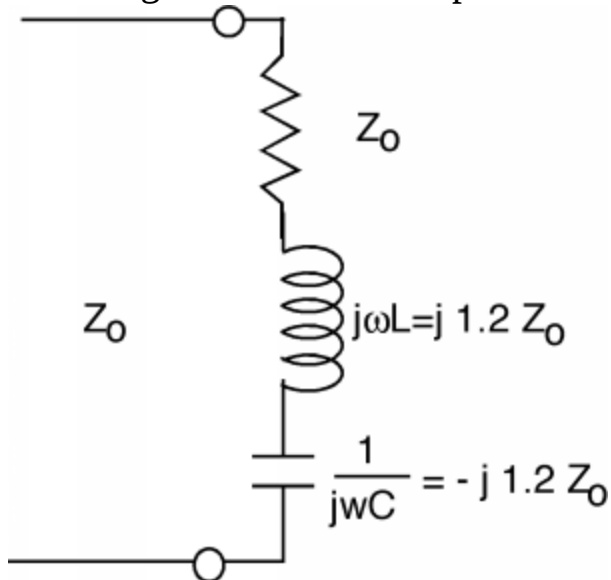
i. Now this is a very



The load impedance

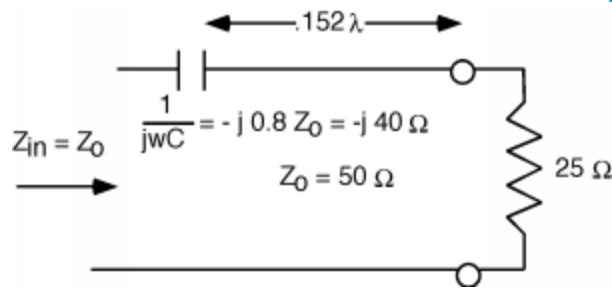
result. Suppose we take the load off the line, and add, in series, an additional capacitor, whose reactance is $\frac{1}{j\omega C} = -j 1.2 Z_0$.

Matching the load with a capacitor

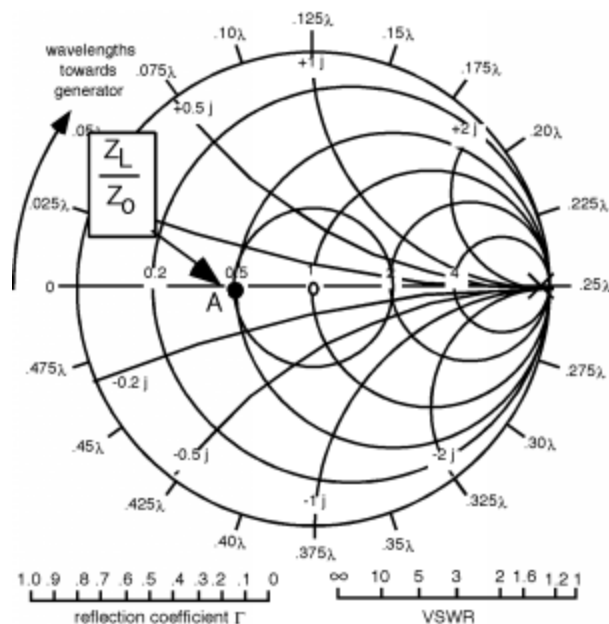


The capacitor and the inductor just cancel each other out (series resonance) and so the apparent load for the line is just Z , the magnitude of the reflection coefficient (Γ) = 0 and the ! All of the energy flowing down the line is coupled to the load resistor, and nothing is reflected back towards the load.

We were lucky that the real part of $\frac{Z}{Z_0}$. If there were not that case, we would not be able to "match" the load to the line, right? Not completely. Let's consider another example. [The next figure](#) shows a line with a Z , terminated with a resistor. Γ —, and we end up with the VSWR circle shown in the [subsequent figure](#).

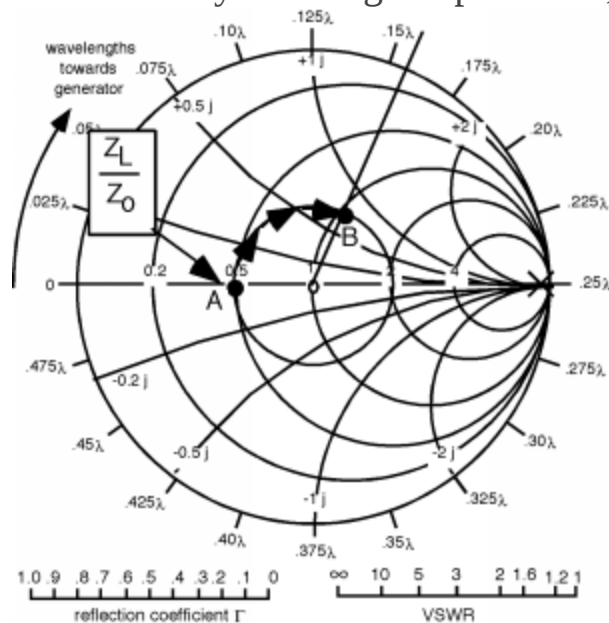


Matching with a series capacitor



Plotting $\frac{Z}{Z_0}$

How could we match this load? We could add another 25Ω in series with the first resistor, but if we want to maximize the power we deliver to the first one, this would not be a very satisfactory approach. Let's move down the line a ways. If we go to point "B", we find that



Moving to the "right spot"

at this spot, $\frac{Z}{Z_0}$ i . Once again we have an impedance with a normalized real part equals 1! How far do we go? It looks like it's a little more than λ . If we add a negative reactance in series with the line at this point, with a normalized value of i , then from that point on back to the generator, the line would "look" like it was terminated with a matched load.

There's one awkward feature to this solution, and that is we have to cut the line to insert the capacitor. It would be a lot easier if we could simply add

something across the line, instead of having to cut it. This is easily done, if we go over into the admittance world.

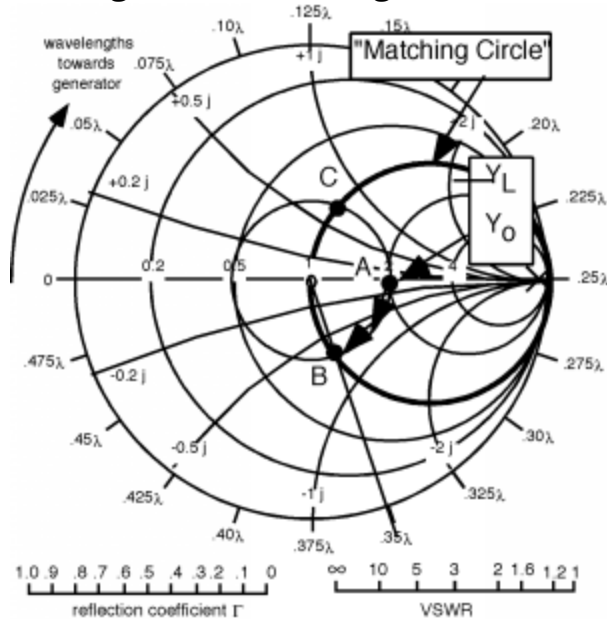
Introduction to Parallel Matching

Let's start with the load. With the same 25Ω resistor for the load, and plot its **admittance** $\frac{Y_L}{Y_0} = 2$. If we start moving away from the load towards the generator, in about 0.10λ we again run into the circle which represents

$$\frac{Y(s)}{Y_0} = 1. \text{ This is such an important circle it has gained its own name,}$$

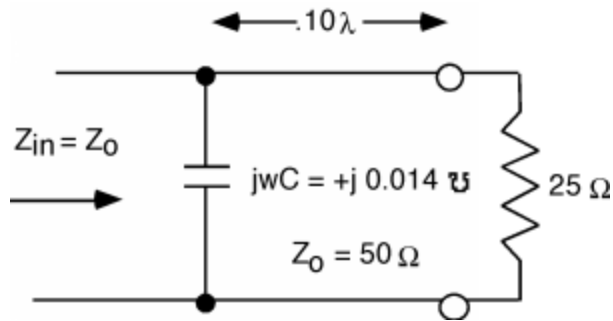
and it is frequently called the **matching circle** [\[link\]](#).

Getting to the Matching Circle



Note that to find out how far we had to move, we had to start at relative position 0.25λ as our zero, or reference location. Point "B" seems to be at about 0.35λ on the scale, and since we started at 0.25λ , the distance is $0.35 - 0.25 = 0.10$. At "B", $\frac{Y_s}{Y_0} = -1.0 + 0.7i$. Thus, if we add a susceptance iB with a value of $i0.014\Omega^{-1}$ we would again match the line. Positive susceptance comes from a capacitor as well, and so [\[link\]](#) shows how we match.

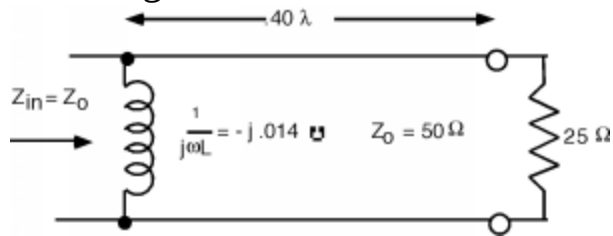
Matching With a Shunt Capacitor



Note that we are not **required** to go to point "B". Any point on the matching circle that we can get to is fair game. Another such point is "C" in [\[link\]](#). This is at a distance of about 0.40λ from the load. At "C",

$$\frac{Y_s}{Y_0} = 1.0 + 0.7i \text{ and so we would put in an inductor, with a susceptance } \frac{1}{i\omega L} = -i0.014\Omega^{-1} \text{ [\[link\]](#)}.$$

Matching With a Shunt Inductor

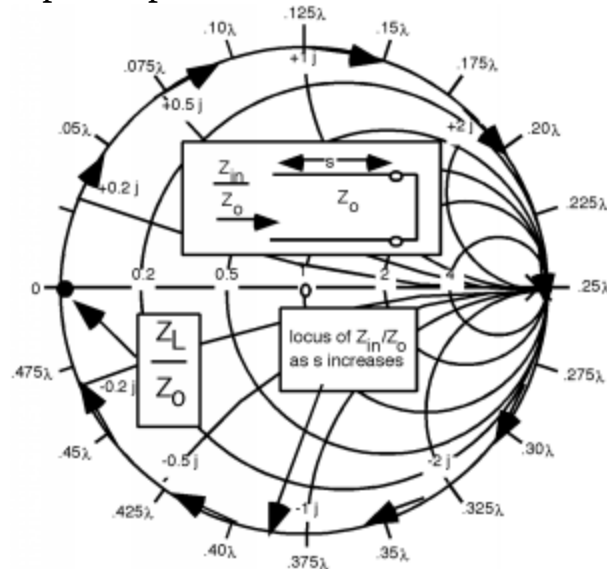


Single Stub Matching

Often, there are reasons why using a discrete inductor or capacitor for matching is not such a good idea. At the high frequencies where matching is important, losses in both L or C mean that you don't get a good match, and most of the time (except for some air-dielectric adjustable capacitors) it is hard to get **just** the value you want.

There is another approach though. A shorted or open transmission line, when viewed at its input looks like a pure reactance or pure susceptance. With a short as a load, the reflection coefficient has unity magnitude $|\Gamma| = 1.0$ and so we move around the very outside of the [Smith Chart](#) as the length of the line increases or decreases, and $\frac{Z_{in}}{Z_0}$ is purely imaginary. When we did the bilinear transformation from the $\frac{Z(s)}{Z_0}$ plane to the $r(s)$ plane, the imaginary axis transformed into the circle of diameter 2, which ended up being the outside circle which defined the Smith Chart.

Input Impedance of a Shorted Line



Another way to see this is to go back to [this equation](#). There we found:

Equation:

$$Z(s) = Z_0 \frac{Z_L + iZ_0 \tan(\beta s)}{Z_0 + iZ_L \tan(\beta s)}$$

With $Z_L = 0$ this reduces to

Equation:

$$Z(s) = iZ_0 \tan(\beta s)$$

Which, of course for various values of s , can take on any value from $i\infty$ to $-(i\infty)$. We don't have to go to Radio Shack© and buy a bunch of different inductor and capacitors. We can just get some transmission line and short it at various places!

Thus, instead of a discrete component, we can use a section of shorted (or open) transmission line instead [\[link\]](#). These matching lines are called **matching stubs**. One of the major advantages here is that with a line which has an adjustable short on the end of it, we can get any reactance we need, simply by adjusting the length of the stub. How this all works will become obvious after we take a look at an example.

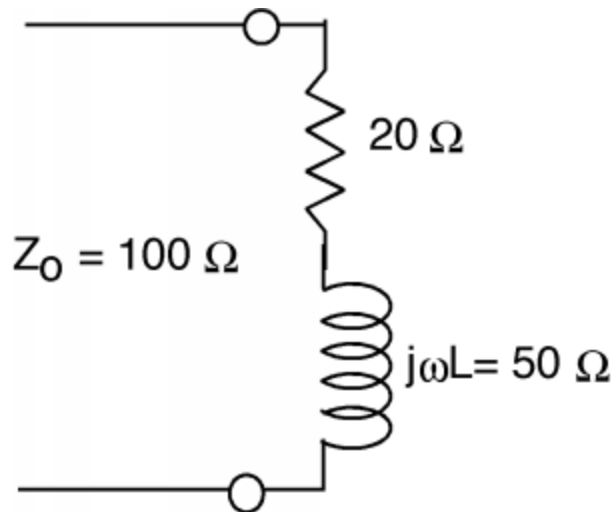
A Shortened Stub



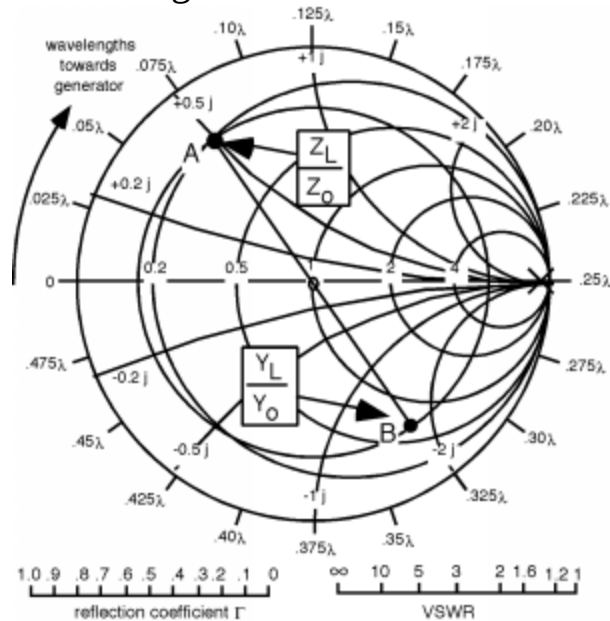
Let's do one. In [\[link\]](#) we can see that, $\frac{Z_L}{Z_0} = 0.2 + 0.5i$, so we mark a point "A" on the Smith Chart. Since we will want to put the tuning or matching stub in shunt across the line, the first thing we will do is convert $\frac{Z_L}{Z_0}$ into a normalized admittance $\frac{Y_L}{Y_0}$ by going 180° around the [Smith Chart](#) to point "B", where $\frac{Y_L}{Y_0} = 0.7 - 1.7i$. Now we rotate around on the constant radius, $r(s)$ circle until we hit the matching circle at point "C". This is shown in [\[link\]](#). At "C", $\frac{Y_S}{Y_0} = 1.0 + 2.0i$. Using a "real" Smith Chart, I get that the distance of rotation is about 0.36λ . Remember, all the way around is $\frac{\lambda}{2}$, so you can very often "eyeball" about how far you have to go, and doing so is a good check on making a stupid math error. If the

distance doesn't look right on the Smith Chart, you probably made a mistake!

Another Load

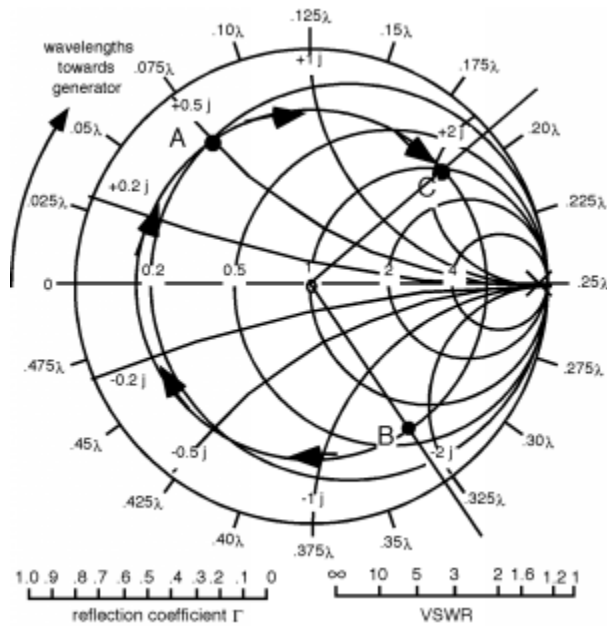


Converting to Normalized Admittance



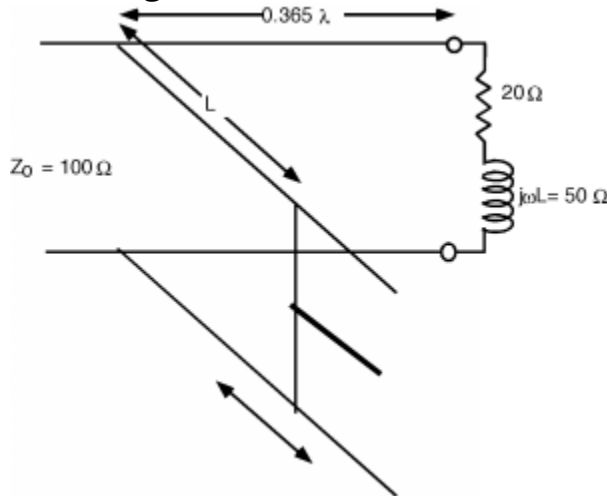
Converting to $\frac{Y_L}{Y_0}$

Moving to the Matching Circle



OK, at this point, the real part of the admittance is unity, so all we have to do is add a stub to cancel out the imaginary part. As mentioned above, the stubs often come with adjustable, or "sliding short" so we can make them whatever length we want [\[link\]](#).

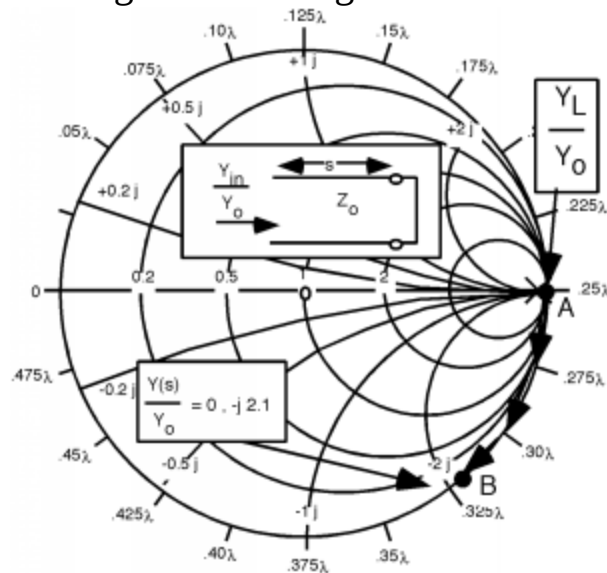
Matching with a Shortened Stub



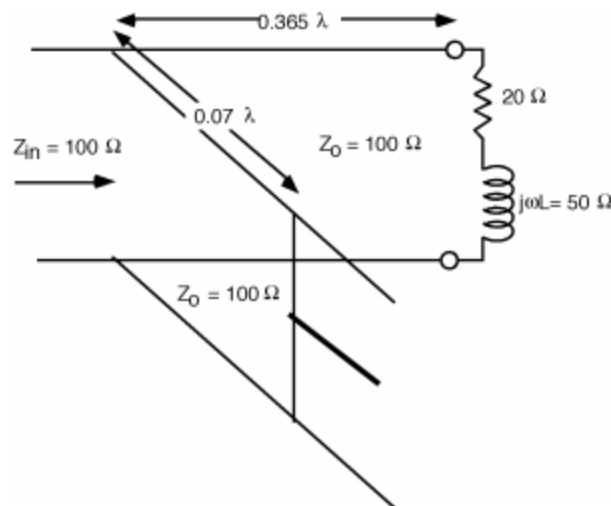
Our task now, is to decide how much to push or pull on the sliding handle on the stub, to get the reactance we want. The hint on what we should do is in [\[link\]](#). The end of the stub is a short circuit. What is the admittance of a short circuit? Answer: $\infty, j\infty$! Where is this on the Smith Chart? Answer: on the outside, on the right hand side on the real axis. Now, if we start at a

short, and start to make the line longer than $s = 0$, what happens to $\frac{Y(s)}{Y_0}$? It moves around on the outside of the Smith Chart. What we need to do is move away from the short until we get $\frac{Y(s)}{Y_0} = -(i2.0)$ and we will know how long the shorted tuning stub should be [\[link\]](#). In going from "A" to "B" we traverse a distance of about 0.07λ and so that is where we should set the position of the sliding short on the stub [\[link\]](#).

Finding the Stub length



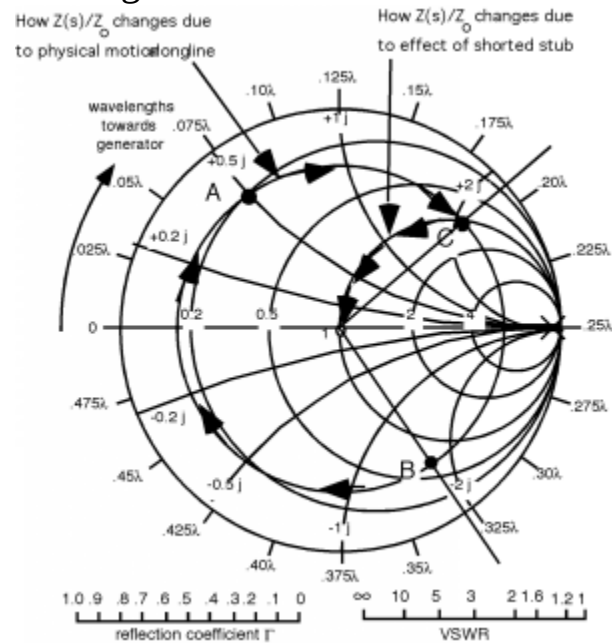
The Matched Line



We sometimes think of the action of the tuning stub as allowing us to move in along the $\frac{Y(s)}{Y_0}$ to get to the center of the Smith Chart, or to a match [\[link\]](#). We are not in this case, physically moving down the line. Rather we

are moving along a **contour of constant real part** because all the stub can do is change the imaginary part of the admittance, it can do nothing to the real part!

Moving With a Stub

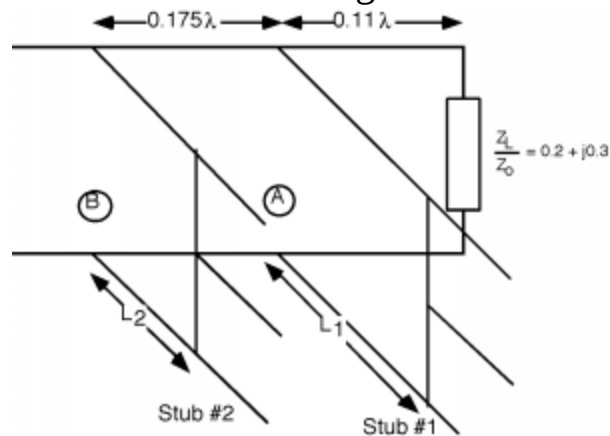


Moving along the
 $\frac{Y(s)}{Y_0} = 1$ circle with a
 stub.

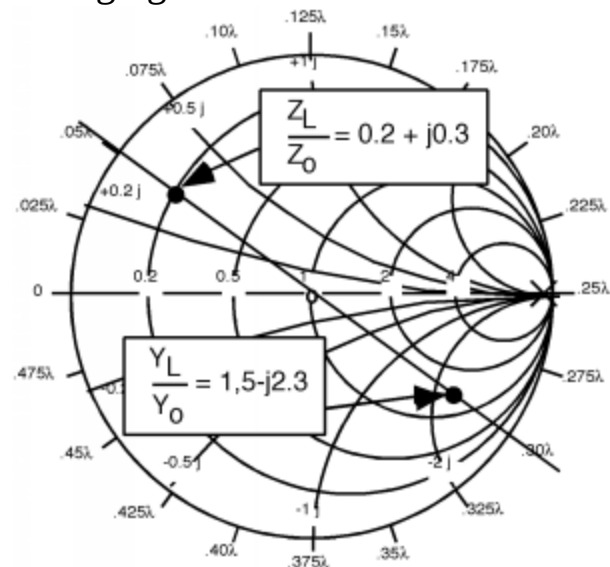
Double Stub Matching

There is one last technique we can look at which is somewhat more flexible than the single stub matching which we just looked at. This is called double stub matching! Suppose we have the following situation, as depicted in [the figure](#). There is a load of $\frac{Z_L}{Z_0} = 0.2 + j0.3$ located at the end of the line, and then some arbitrary distance away () an adjustable stub. Another (arbitrary) from the first stub, there is a **second** one. Let's plot $\frac{Z_L}{Z_0}$ on the [Smith Chart](#), and then, since the stubs are in shunt across the line, switch to admittance, and find $\frac{Y_L}{Y_0}$. It is easy to see that $\frac{Y_L}{Y_0} = 1.5 - j2.3$.

Double Stub Matching Problem

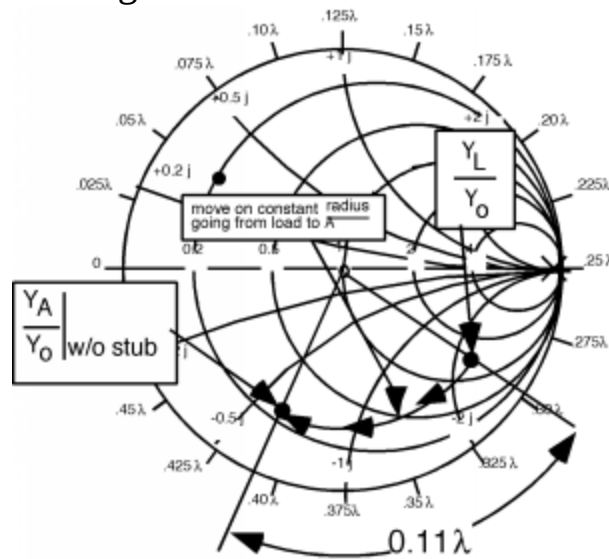


Changing the Load to an Admittance

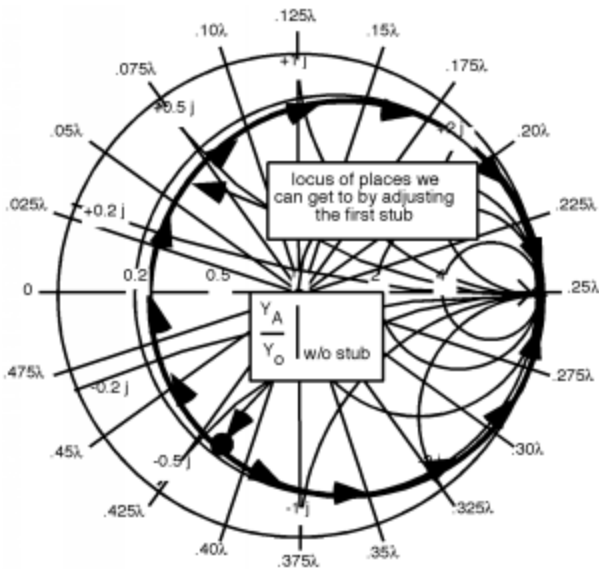


The first thing we might as well do is move down to the first stub, and see what admittance we have there [\[link\]](#). We go from the load, to the first stub by rotating on a **circle of constant radius**(constant r) since all we are doing is going from one place on the line to another. If we call the location on the line of the first stub "A", then we can see that — .

Moving From the Load to the First Stub



Now, what can the first stub accomplish? A shorted stub can create any **imaginary** admittance we want, but can not change the real part of the admittance. Thus, by adjusting the first stub, we can move around on a circle of **constant real part** r , and have any imaginary part we want. This is shown schematically [here](#).
Possible Effects of the First Stub



Now, where do we want to go? Well, we would like to end up someplace so that, after we have moved from A to B on the line (gone from the first stub to the second), we are on the matching circle. If this were so, then, since we are on the matching circle, we could use the second stub to match the whole line and we would be done.

This is tricky now, so you have to pay attention and think. If I want to find a place which, when moved from A to B, ends up on the matching circle, then what I should do is take the matching circle and move it from B to A. That is, if I rotate the matching circle around **towards the load**, then any place on that rotated matching circle is guaranteed to end up on the **real** matching circle, when we go back towards the generator.

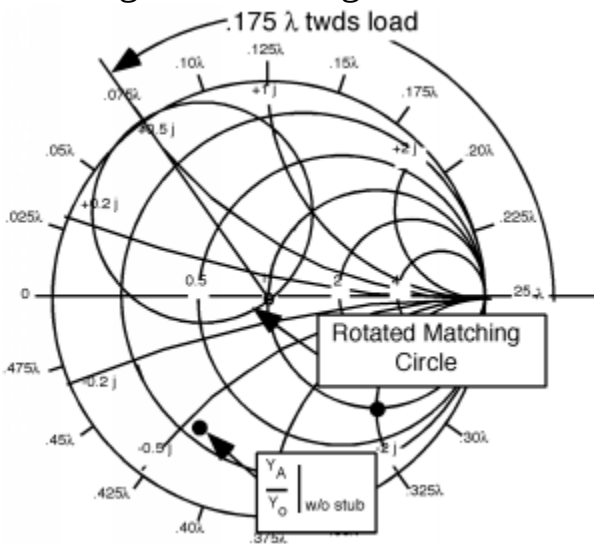
OK, so here's what we do. First, we rotate the matching circle 0.175 around towards the load (go counterclockwise) [\[link\]](#). Now what we have to do is somehow get from — without stub to someplace on the rotated matching circle. The only way we can do this is to change the imaginary part of with the stub. Suppose we move as shown in [\[link\]](#). In going from — without stub to — with stub we have changed the imaginary part from

to , thus we have **added** to the imaginary part of —.

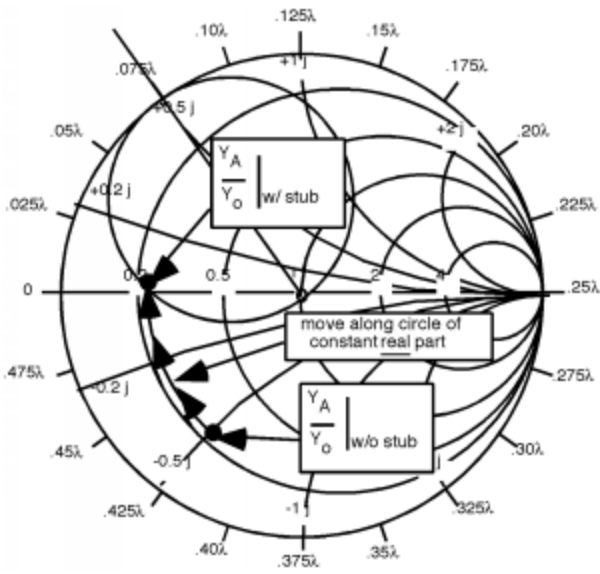
Thus using our standard method for finding the length of the first stub, we start at , (the short at the end of the stub) and go around the outside

of the [Smith Chart](#) until we find $1 + j1$. To get from one place to the next we went 0.175λ and so the length of the first stub, 0.175λ should be 0.175λ . Now we are at $1 + j0$ with stub. The next thing we have to do is to rotate another 0.175λ **towards the generator** so that we can get to stub B. As we do this rotation, we again stay on a circle of constant **radius**, because now we are moving down the transmission line **not** adding reactance by using a stub! This rotation is **guaranteed** to end us up on the matching circle because **every** point on the rotated circle (the one we start from) is exactly 0.175λ towards the load **from** the matching circle. As shown [here](#), we are now at the point $1 + j0$ without stub 0.175λ . Thus we need to adjust the length of the second stub to give us 0.175λ of reactance, so we can move (along a circle of **constant real part** = 1.0) into the center of the [Smith Chart](#). We have to find the length for the second stub, but that is now easy! ([link](#))

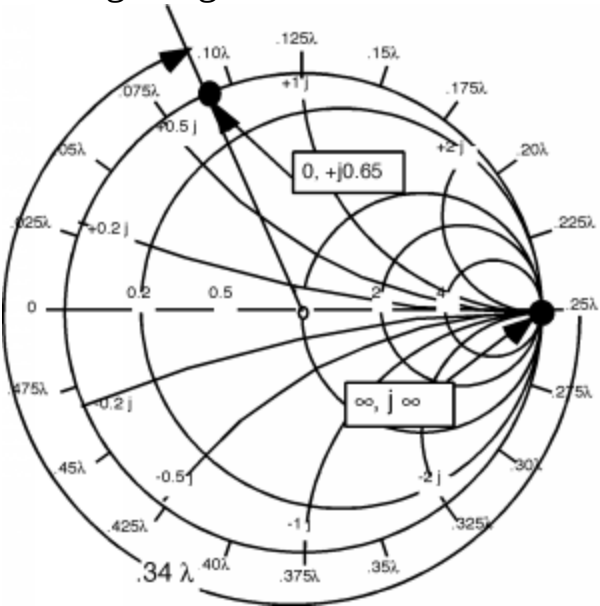
Rotating the Matching Circle



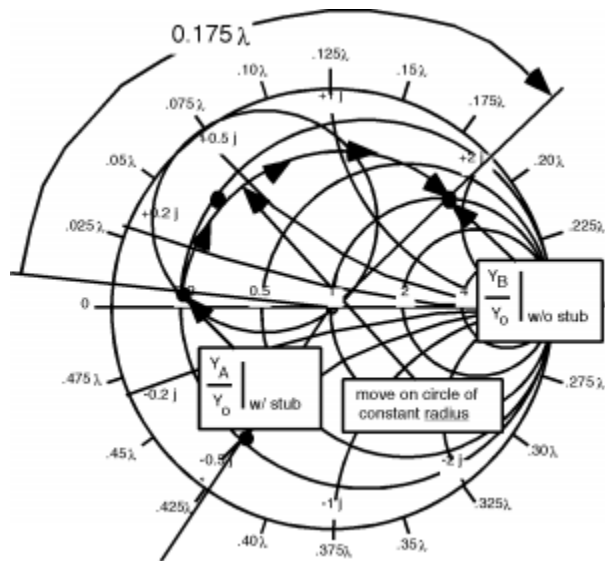
Moving to Rotated Matching Circle



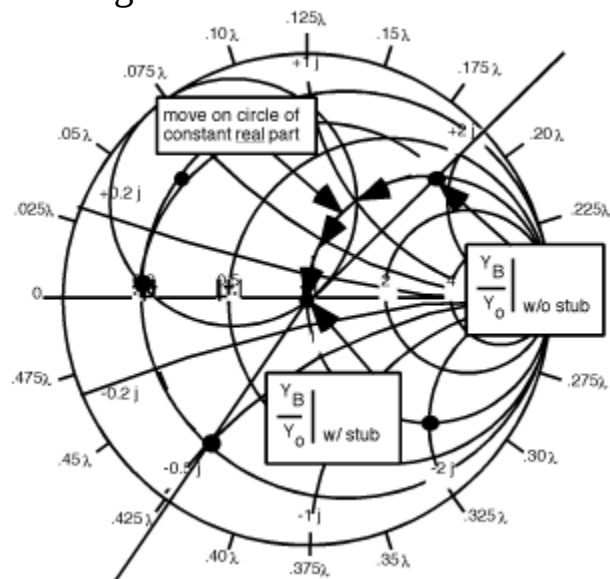
Finding Length of the First Stub



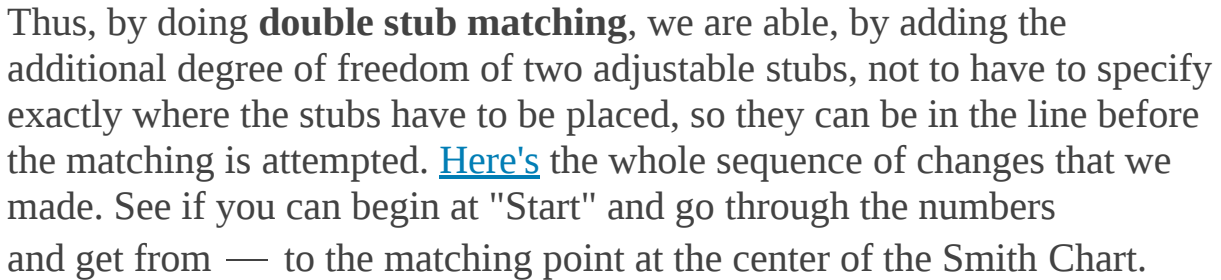
Moving Down the Second Stub



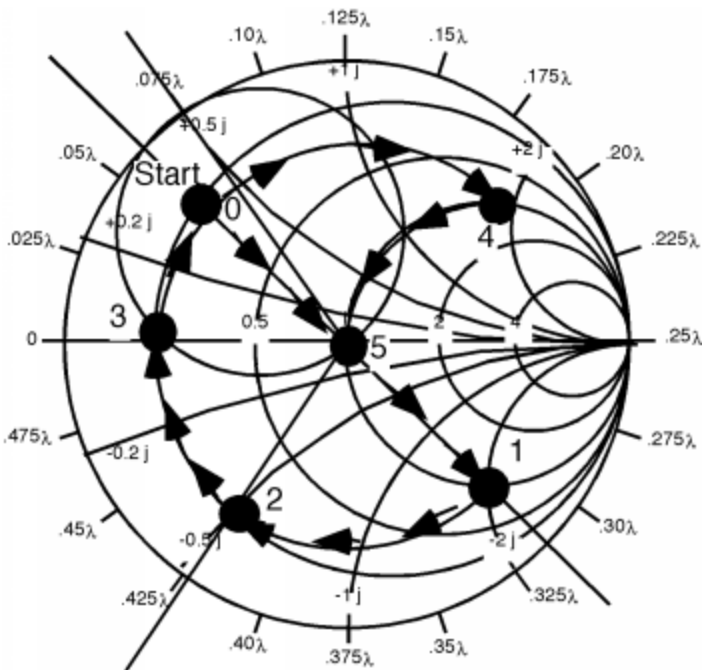
Making the Match



Finding the Length of the Second Stub

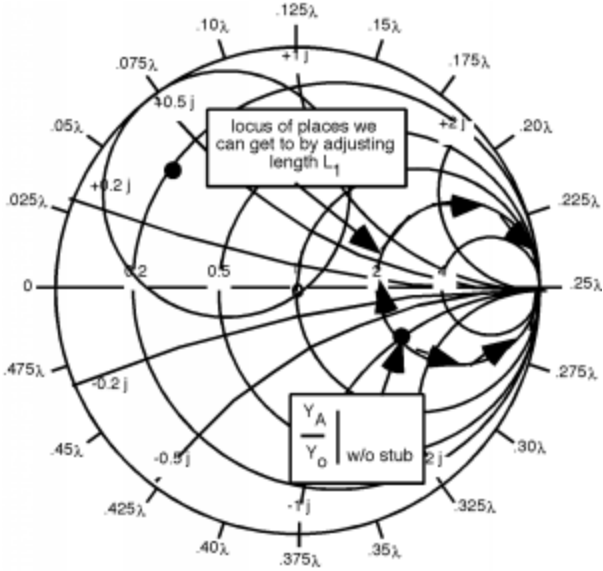


Double Stub Matching All Put Together!



There's just one little problem. What if — without stub had ended up as shown in [here](#). We are on the — circle. No matter how hard I try, and no matter where I set all I can do is spin around on the little circle as [shown](#), and I will never end up on the rotated matching circle, and I won't be able to make a match! Well, if I add a **third stub**...I'll let you work it out!

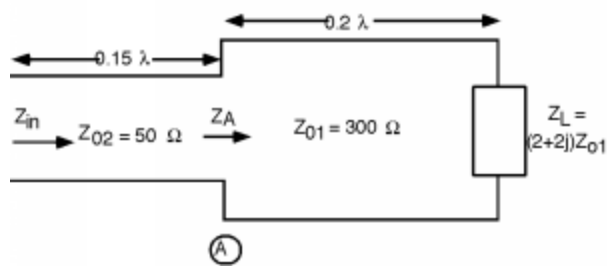
A Situation That Doesn't Work



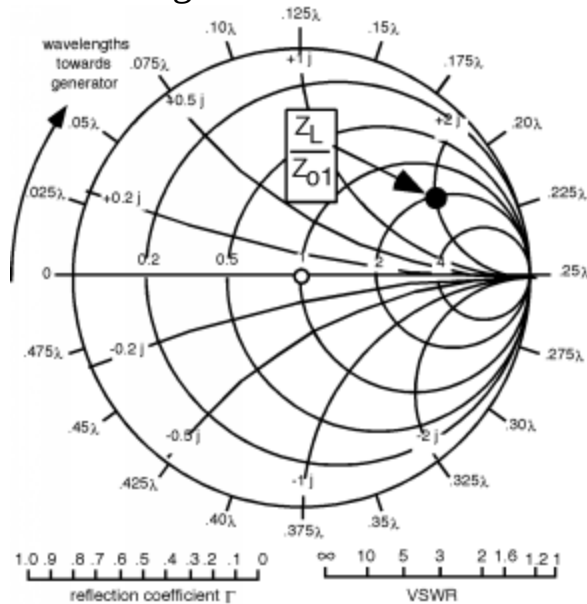
Odds and Ends

Just a few odds and ends. Consider the [following](#) which is called a "cascaded line" problem. These are problems where we have two different transmission lines, with different characteristic impedances. Since we will give all of the distances in wavelengths, λ , we will assume that the λ we are talking about is the appropriate one for the line involved. If the phase velocities on the two lines is the same, then the physical lengths would correspond as well. The approach is relatively straight-forward. First let's plot $\frac{Z_L}{Z_0}$ on the [Smith Chart](#). Then we have to rotate 0.2λ so that we can find $\frac{Z_A}{Z_{01}}$, the normalized impedance at point A, the junction between the two lines [\[link\]](#).

Cascaded Line

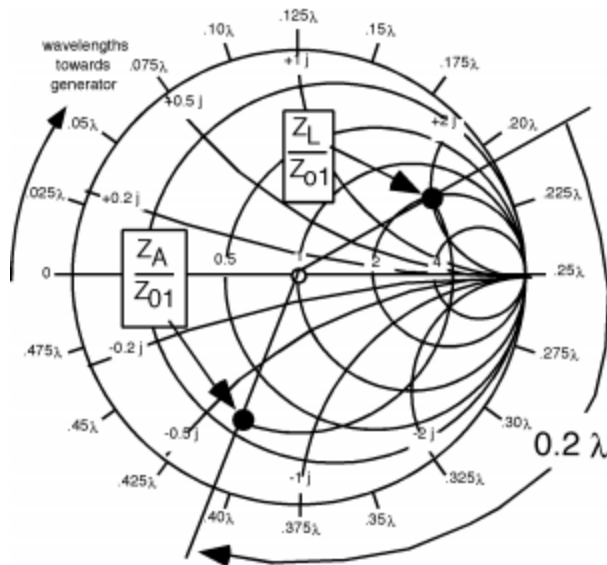


Smith Diagram

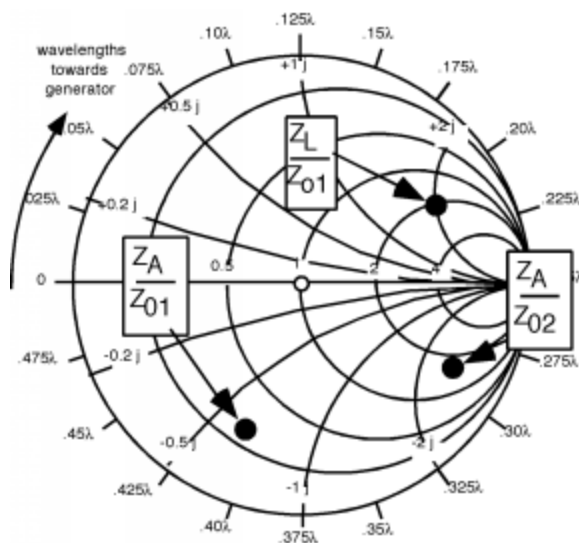


Thus, we find $\frac{Z_A}{Z_{01}} = 0.32 + 0.6i$. Now we have to **renormalize** the impedance so we can move to the line with the new impedance Z_{02} . Since $Z_{01} = 300(\Omega)$, $Z_A = 96 + -180i$. This is the load for the second length of line, so let's find $\frac{Z_A}{Z_{02}}$, which is easily found to be $1.9 + -3.6i$, so this can be plotted on the [Smith Chart](#). Now we have to rotate around another 0.15λ so that we can find $\frac{Z_{in}}{Z_{02}}$. This appear to have a value of about $0.15 + -0.45i$, so $Z_{in} = 7.5 + -22.5i\Omega$ [\[link\]](#).

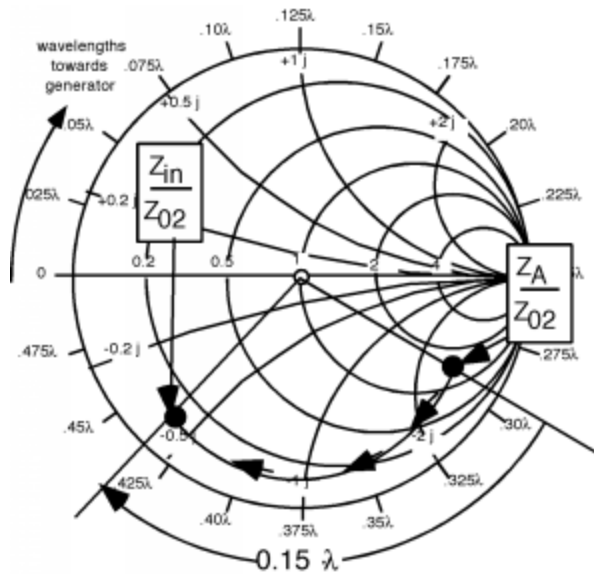
Towards the Generator



More Smith Charts



Even More Smith Charts

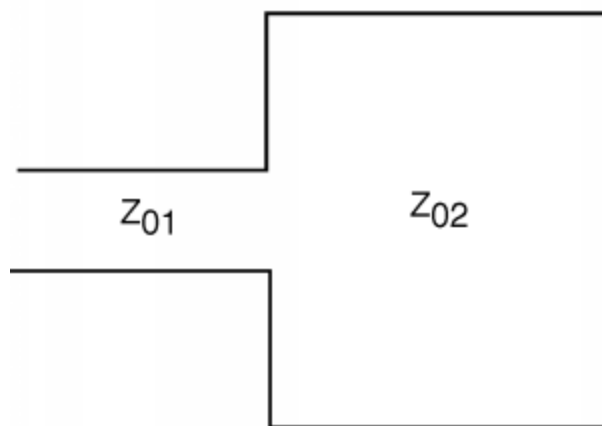


There is one application of the cascaded line problem that is used quite a bit in practice. Consider the following: We assume that we have a matched line with impedance Z_{02} and we connect it to another line whose impedance is Z_{01} [\[link\]](#). If we connect the two of them together directly, we will have a reflection coefficient at the junction given by

Equation:

$$\Gamma = \frac{Z_{02} - Z_{01}}{Z_{02} + Z_{01}}$$

Simplified Cascaded Line

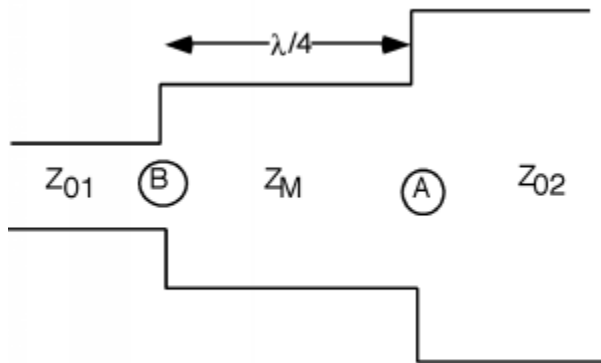


Now let's imagine that we have inserted a section of line with length $l = \frac{\lambda}{4}$ and impedance Z_m [\[link\]](#). At point A, the junction between the first line and the matching section, we can find the normalized impedance as

Equation:

$$\frac{Z_A}{Z_M} = \frac{Z_{02}}{Z_m}$$

Another Cascaded Line



We take this impedance and rotate around on the Smith Chart $\frac{\lambda}{4}$ to find $\frac{Z_B}{Z_M}$

Equation:

$$\frac{Z_B}{Z_M} = \text{mfrac}$$

where we have taken advantage of the fact that when we go half way around the Smith Chart, the impedance we get is just the inverse of what we had originally (half way around turns $r(s)$ into $-r(s)$).

Thus

Equation:

$$Z_B = \frac{Z_m^2}{Z_{02}}$$

If we want to have a match for line with impedance Z_{01} , then Z_B should equal Z_{01} and hence:

Equation:

$$\begin{aligned} Z_B &= Z_{01} \\ &= \frac{Z_m^2}{Z_{02}} \end{aligned}$$

or

Equation:

$$Z_m = \sqrt{Z_{01} Z_{02}}$$

This piece of line is called a **quarter wave matching section** and is a convenient way to connect two lines of different impedance.